PSYCHOPHARMACOLOGIC DRUGS ADVISORY COMMITTEE

FDA White Oak Campus

10903 New Hampshire Ave.

Bldg. 31, Conference Center

Silver Spring, MD

January 12, 2016

PROBUPHINE

(buprenorphine hydrochloride subdermal implant)

for maintenance treatment of opioid dependence

DISCLAIMER STATEMENT

The attached package contains background information prepared by the Food and Drug Administration (FDA) for the panel members of the advisory committee. The FDA background package often contains assessments and/or conclusions and recommendations written by individual FDA reviewers. Such conclusions and recommendations do not necessarily represent the final position of the individual reviewers, nor do they necessarily represent the final position of the Review Division or Office. We have brought NDA 204442, PROBUPHINE (buprenorphine hydrochloride) subdermal implant, submitted by Titan Pharmaceuticals, Inc., to this Advisory Committee in order to gain the Committee's insights and opinions, and the background package may not include all issues relevant to the final regulatory recommendation and instead is intended to focus on issues identified by the Agency for discussion by the advisory committee. The FDA will not issue a final determination on the issues at hand until input from the advisory committee process has been considered and all reviews have been finalized. The final determination may be affected by issues not discussed at the advisory committee meeting.

FOOD AND DRUG ADMINISTRATION

Center for Drug Evaluation and Research

Meeting of the Psychopharmacologic Drugs Advisory Committee

PROBUPHINE (buprenorphine hydrochloride subdermal implant)
for maintenance treatment of opioid dependence

January 12, 2016

BRIEFING MATERIALS

Table of Contents

Food and Drug Administration
CENTER FOR DRUG EVALUATION AND RESEARCH
Division of Anesthesia, Analgesia, and Addiction Products

MEMORANDUM

DATE: December 15, 2015

FROM: Sharon Hertz, Division Director
Division of Anesthesia, Analgesia, and Addiction Products

TO: Chair, Members, and Invited Guests, Psychopharmacologic Drugs Advisory Committee (PDAC)

RE: Overview of the January 12, 2016 PDAC Meeting to discuss NDA 204442 for Probuphine (buprenorphine subdermal implant) for treatment of opioid dependence

At this meeting of the Psychopharmacologic Drugs Advisory Committee, we will be discussing a new drug application (NDA) 204442, for Probuphine (buprenorphine subdermal implant) submitted by Titan Pharmaceuticals, Inc., for the treatment of opioid dependence.

As an implantable formulation, Probuphine is intended to be difficult to divert or abuse, and less likely than sublingual buprenorphine formulations to be accidentally ingested by small children. It also offers the possibility of improved adherence to treatment and improved patient convenience. However, because it must be surgically inserted and removed, it requires specific training.

During this meeting, representatives from the Agency and the Applicant will present background on the initial submission of this application which led the Agency to conclude that the dose provided was too low to be effective for a broad population of patients, and the Applicant's subsequent decision to pursue approval for treatment of patients who are already stable on low-to-moderate doses of sublingual buprenorphine. The efficacy data from a new clinical trial, Study PRO-814, conducted by the Applicant to assess Probuphine in the treatment of opioid-dependent patients already stable on buprenorphine will be presented, particularly in the context of the novel study design with respect to population, design, and analytic approach. The safety data from the clinical program will be reviewed including data from the original clinical trials. Because the systemic safety of buprenorphine has been characterized already, the safety

presentation will emphasize information about adverse events associated with the surgical insertion and removal procedures. The Applicant's proposed Risk Evaluation and Mitigation Strategy (REMS) will also be discussed.

Following these presentations, you will be asked to assess these findings and to discuss the following aspects of the application:

1. The factors that define a stable patient for the purposes of patient selection, and whether the Applicant has successfully identified a patient population that can benefit from Probuphine.

2. The factors (e.g., extent and timing of use of rescue, nature of missing urine toxicology data) that should be included in classifying responders.

3. Whether the Applicant has provided adequate evidence that Probuphine is effective in stable patients, taking into consideration the data on rescue use, the nature of missing information, and the differences between the assumptions used to set the non-inferiority margin and the observed results.

4. How the need for occasional supplemental doses will translate to clinical practice, and what guidance should be provided in labeling for clinicians to identify patients who are not adequately treated with Probuphine. In particular, if prescriptions for as-needed sublingual buprenorphine are anticipated to be a routine practice, how these prescriptions will impact the product's ability to mitigate misuse, abuse, and accidental pediatric exposure.

5. Whether the Risk Evaluation and Mitigation Strategy (REMS) proposed by the Applicant, which consists of restricted distribution and a training/certification program for providers who will insert and remove the product, is adequate to address the risks of potential procedure complications.

The Division and the Agency are grateful to the members of the committee and our invited guests for taking time from your busy schedules to participate in this important meeting. Thank you in advance for your advice, which will aid us in making the most informed and appropriate decision possible.

Efficacy and Safety Background

LIST OF TABLES

LIST OF FIGURES

1 Executive Summary

Probuphine is a rod-shaped implant designed to provide sustained delivery of buprenorphine, a partial agonist at the μ-opiate receptor, for up to six months when 4 rods[1] are inserted subdermally. Probuphine is intended as a maintenance treatment for opioid-dependent patients who are clinically stable on a low dose of sublingual buprenorphine (equivalent to 8 mg/day or less of buprenorphine as Suboxone tablet[2]).

Titan[3], the Applicant, has provided efficacy data from a single, double-blind, double-dummy, active-controlled trial. The study design includes a number of novel features not seen in prior studies of drugs used to treat opioid dependence.
These include:

- Enrollment of clinically-stable patients
- Infrequent verification of abstinence from illicit drug use, consistent with the frequency of clinical monitoring of stable patients
- Use of an active-control design with the objective of demonstrating non-inferiority[4] of Probuphine to an active control

The Committee will be asked to consider whether the data from the clinical trial provide substantial evidence of effectiveness of Probuphine for the maintenance treatment of opioid dependence in a subset of patients with opioid dependence.

The Applicant's submission includes safety data from 309 unique patients who were treated with Probuphine. The overall safety experience is consistent with the known safety profile of buprenorphine. However, the product presents a novel safety concern among products used to treat opioid dependence associated with the surgical insertion and subsequent need for removal of the implanted rods. It is similar in many respects to Norplant, an implantable, progestin-releasing contraceptive which is no longer marketed in the US.

Despite the fact that insertion and removal of Norplant were performed by providers trained in surgery, the product's safety experience identified the potential for various insertion and removal-related complications, some of them with disabling consequences. Similar difficulties may be anticipated with Probuphine, perhaps further complicated by

[1] The terms "implant" and "rod" are used interchangeably throughout this document.
[2] Subutex and Suboxone tablets are no longer marketed by the manufacturer. Equivalent doses may include generic buprenorphine tablets at a dose of 8 mg buprenorphine; generic buprenorphine/naloxone tablets 8 mg/2 mg buprenorphine/naloxone; Zubsolv tablets at a dose of 5.7 mg/0.71 buprenorphine/naloxone; Bunavail buccal film at a dose of 4.2 mg/0.7 mg buprenorphine/naloxone. Suboxone Film, 8 mg/2 mg buprenorphine/naloxone delivers a somewhat higher exposure to buprenorphine than Suboxone tablets at the same dose.
[3] Braeburn is the authorized agent of the Applicant.
[4] The goal of a non-inferiority study is to show that a test drug is not unacceptably worse than an active control. This is accomplished by showing that lower bound of the 95% confidence interval for the difference between the test drug and the active control is greater than some pre-specified margin. Non-inferiority will be discussed in greater depth in Section 5.3.1

the population of both prescribers who generally lack surgical training and patients who present more heterogeneity in age, sex, and health status compared to the population being treated with Norplant. The Applicant has proposed a training program for providers, and a closed distribution system to ensure the insertion and removal procedures are performed only by trained providers, to address this concern. The committee will be asked to address whether this concern, or any additional safety concerns, have been adequately addressed by the existing safety data, and can be adequately managed under the proposed Risk Evaluation and Mitigation Strategy (REMS). Finally, the committee will be asked whether the efficacy data are sufficient to outweigh the risks associated with this novel product.

2 Introduction and Background

Buprenorphine is a partial agonist at the μ-opiate receptor. A parenteral formulation of buprenorphine was approved in 1981 for the treatment of pain, and two sublingual tablet formulations were approved in 2002 for the treatment of opioid dependence[5]. Three other transmucosal formulations have subsequently been approved. Approximately 10.7 million prescriptions were dispensed from outpatient retail pharmacies and approximately 1 million patients received a dispensed prescription for buprenorphine tablets or films during 2012.[6]

Buprenorphine was developed as a treatment for opioid dependence because some of its pharmacological properties suggested it could serve as a safer alternative to methadone, a full agonist at the μ-opioid receptor. First, buprenorphine had been shown to have a ceiling effect for respiratory depression, suggesting that it would be "impossible to overdose" on buprenorphine. Second, initial clinical evaluations of buprenorphine's ability to produce physical dependence led to the conclusion that physical dependence to buprenorphine, if it developed, was associated with a mild withdrawal syndrome. Third, it was expected to have limited attractiveness as a drug of abuse relative to full agonists.[7]

Buprenorphine was expected to have limited abuse potential for two reasons. First, due to its partial agonist properties, the euphorigenic effects of buprenorphine were understood to reach a "ceiling" at moderate doses, beyond which increasing doses of the drug do not produce the increased effect that would result from full opioid agonists. Second, when a partial agonist displaces a full agonist at the receptor, the relative reduction in receptor activation can produce withdrawal effects. Individuals dependent on full agonists may therefore experience sudden and severe symptoms of withdrawal if they use buprenorphine. These features were expected to limit its attractiveness as a drug of abuse for patients and for illicit use.

In addition to the improved safety profile, at sufficiently high doses, buprenorphine blocks full opioid full agonists from achieving their full effects, deterring abuse of other opioids by buprenorphine-maintained patients.

---

[5] Subutex, buprenorphine sublingual tablets (Reckitt Benckiser NDA 20732) and Suboxone, buprenorphine/naloxone sublingual tablets (Reckitt Benckiser NDA 20733). Naloxone is intended to further deter abuse by the intravenous route by precipitating withdrawal if the product is injected by persons dependent on full agonists.

[6] Additionally, in a Drug Enforcement Administration (DEA) report on Buprenorphine, the DEA reports that IMS Health™ National Prescription Audit Plus indicates from January to March 2013, 2.5 million buprenorphine prescriptions were dispensed. The DEA Report on Buprenorphine is available at: http://www.deadiversion.usdoj.gov/drug_chem_info/buprenorphine.pdf

[7] Many of these beliefs have subsequently been found to have been erroneous, or at least overstated, but these were the generally-held views about buprenorphine's pharmacology at the time it was being developed.

Unfortunately, despite these features, buprenorphine sublingual products have been increasingly identified in the illicit drug market, and it is known that they are diverted, abused, and misused. Additionally, they have been implicated in a number of cases of accidental poisonings of small children. Therefore, a depot injection or an implantable product which would be difficult to divert or abuse, and would be less likely to be accidentally ingested by small children, offers potential advantages. In addition, if a depot or implantable product provided a sufficient plasma level of buprenorphine to block the effects of exogenous opioids, the nature of the product would enforce compliance so that patients could not periodically discontinue use to allow the blocking effect to dissipate in order to experience the effects of their opioids of choice.

The recommended dose of sublingual buprenorphine is in the range of 12 mg to 16 mg daily. Pharmacokinetic comparisons of Probuphine to sublingual buprenorphine demonstrate that the relative bioavailability of four Probuphine implants (320 mg total buprenorphine) based on the mean AUC_{0-24} values at steady state compared with sublingual buprenorphine (16 mg once daily) is 31.3%. The trough concentrations of buprenorphine at steady-state obtained with Probuphine were approximately 0.72 to 0.83 ng/mL, approximately half the trough concentrations observed with 16 mg daily of sublingual buprenorphine at steady state (1.6 ± 0.6 ng/mL).

2.1 FDA-Approved Products for the Treatment of Opioid Dependence

Other approved products for the treatment of opioid dependence include buprenorphine sublingual formulations, including buprenorphine in combination with naloxone; methadone and levomethadyl acetate (LAAM, no longer marketed), both of which are full agonist treatments; and naltrexone (oral and depot formulations), an opioid antagonist. Treatment of addiction with methadone is limited to closely-regulated Opioid Treatment Programs (OTP), which may limit access to treatment. Buprenorphine treatment may be prescribed by specially-qualified physicians in office practice settings. (See Appendix A.)

2.2 Clinical Development of Probuphine

2.2.1 Original NDA Submission

The original development program undertaken by the Applicant included two placebo-controlled trials that enrolled new entrants to buprenorphine treatment. The appropriate approach to take in designing clinical trials to evaluate treatments for opioid addiction continues to evolve and so there is no standard approach to the clinical trial design of studies that evaluate treatment of opioid dependence. The Applicant conducted the original development program for this indication with advice from the Agency on the trial design and analytic approach.

The Applicant initially envisioned Probuphine as a product which could be provided to patients at the outset of their treatment—after just a few days of titration on a sublingual formulation. To support this indication, the Applicant was asked to provide evidence from replicated trials showing that Probuphine was appropriate treatment for patients who might not yet be stabilized on buprenorphine. The results of these studies, although meeting the pre-specified endpoints, pointed to a conclusion that the dose provided was too low to provide effective treatment for patients new to buprenorphine treatment.

Ultimately, the application was not approved. Appendix B provides, in detail, information about the original clinical trials and the Division's interpretation of the results.

2.2.2 Post-Action Discussions and Development Activities

As the Division concluded, on review of the data, that the dose of Probuphine was too low to be effective, the Applicant was encouraged to study a higher dose of Probuphine. Although the Applicant disagreed with the Division's conclusions regarding the efficacy findings, they did acknowledge that four Probuphine implants yield buprenorphine concentrations similar to those observed with 4 to 8 mg sublingual buprenorphine based on average exposure (e.g., mean AUC values) or concentration. It was noted that, when the study results were discussed at a meeting of the Psychiatric Drugs Advisory Committee on March 21, 2013, experts on the panel commented that there could be a subset of long-term patients stable on lower doses of buprenorphine who could benefit from the product. In accordance with this finding, the Applicant proposed a revised indication for Probuphine of *for the treatment of patients stabilized on sublingual buprenorphine at doses of 8 mg or less*. The Division agreed that, with adequate support, the revised indication may be suitable for a subset of patients given the public health benefit that Probuphine could potentially offer related to decreased misuse, abuse, and accidental pediatric exposure.

Ultimately, to support the revised indication, Study PRO-814 was designed and conducted by the Applicant to assess the efficacy of Probuphine in this new population. Certain aspects of the study design were novel. Customarily, studies of drugs to treat opioid addiction have featured frequent visits for collection of urine toxicology tests to ascertain abstinence from illicit drug use. However, because stable patients already in established buprenorphine treatment would not ordinarily be seen thrice-weekly, or even weekly, the burden on participants was seen as a barrier to participation and likely to lead to discontinuations and missing data. Additionally, there was discomfort with the idea of any design that withdrew stable patients from an effective treatment, putting them at risk for relapse which might not be readily reversed. Therefore, the Division and the Applicant jointly agreed that a double-blind, double-dummy non-inferiority study with sublingual buprenorphine in patients already stable on buprenorphine treatment could be conducted. Although it might be argued that a passive-compliance formulation such as Probuphine should be superior to a formulation that relies upon patients to adhere to a medication regimen, the regulations do not require that a new medication be shown to be superior to an approved medication.

Customarily, non-inferiority studies require that a treatment have a known and consistent effect in order to support the assumptions used to choose the non-inferiority margin. Therefore, historically, the Division has been reluctant to agree to non-inferiority designs for trials of drugs intended to treat opioid dependence because of the lack of consistent information about the expected response rate, related to the heterogeneity of response definitions, study designs, populations, and treatments. However, some flexibility was deemed appropriate because the Division recognized the potential public health benefit of an implantable formulation of buprenorphine in light of a growing problem of misuse, abuse, and accidental exposure of buprenorphine. The Division encouraged the Applicant to seek various sources of information about the expected rate of non-relapse in stable, successfully-treated patients who continue on buprenorphine over a six-month period.

Because of these uncertainties regarding a non-inferiority design in this setting, the Division informed the Applicant that the determination regarding whether a study meeting the proposed primary endpoint, augmented by the secondary endpoints, would provide the adequate evidence necessary to support a label for "the treatment of patients stabilized on sublingual buprenorphine at doses 8 mg or less with four Probuphine subdermal implants" would be a matter for review. The Applicant was further informed that the review would quantitatively and qualitatively assess the analysis of the primary endpoint and the clinical meaningfulness of the trial findings to make such a determination.

In Section 5.1, below, the sources of information used to establish the responder definition, expected responder rate, and non-inferiority margin are discussed.

2.3 Safety Concerns Related to Surgically Implantable Drugs

The Agency's previous experience with surgically implantable products, specifically contraceptive implants, was used to identify potential concerns that could arise in the use of Probuphine, as well as upon the experience in the development program itself.

Implantable methods of contraception consist of devices that can be placed subcutaneously to provide long-acting, readily-reversible contraception. Four iterations of contraceptive implants have been approved for marketing in the United States, with each new generation featuring product designs aimed at improving tolerability. Norplant, the first generation of contraceptive implant, consisted of six levonorgestrel-containing capsules and was approved in 1990. Subsequent versions of implants include Jadelle (a two-capsule, levonorgestrel-containing implant), Implanon (a single-capsule, etonogestrel-containing implant), and Nexplanon (similar to Implanon, but is radio-opaque and detectable by X-ray). Currently, only Nexplanon is marketed in the U.S.

While implantable contraceptive methods are generally well-tolerated, notable procedure-related adverse events include pain, infection, numbness, and scarring at the implant site. Complications such as bleeding or hematoma have also been reported. The most significant safety concerns include injuries related to damage of the ulnar or medial cutaneous nerve, which have resulted in permanent disability.

Notably, implantable contraceptive products are inserted and removed by obstetrician/gynecologists, who are surgically trained. Their medical offices are suitably equipped for the performance of minor surgical procedures; they have access to imaging modalities (such as ultrasound) for localizing implants that cannot be palpated, and to operating suites if a more extensive surgical procedure is required to manage a complication. In contrast, buprenorphine treatment is currently provided by physicians who may not have suitable training and may not practice in suitable environments to permit them to perform the insertion or removal procedures, or to manage complications.

Drug utilization data in 2012 indicated that 32% of prescriptions for buprenorphine/naloxone sublingual tablets are written by physicians whose specialty is identified as General Practitioner/Family Medicine/Doctor of Osteopathy. While some of these individuals may perform minor surgical procedures, others may not be prepared to do so. Fully 22% of prescriptions were written by psychiatrists, who are not routinely trained to perform surgical procedures, and whose office environments are not generally suitable for managing procedural complications associated with insertion and removal of the implants. Internists wrote 16% of prescriptions, while only a very small proportion of prescriptions are written by physicians whose specialties involve surgical training.

3 Clinical Pharmacology

The two clinical pharmacology studies [TTP-400-02-01 and PRO-810] demonstrated that steady-state buprenorphine exposures obtained with four implants (80 mg of buprenorphine each, 320-mg total) were approximately 0.72 to 0.83 ng/mL, which are approximately half the trough concentrations observed with 16 mg/day SL buprenorphine at steady state (1.6 ± 0.6 ng/mL). The relative bioavailability of Probuphine implants (320 mg total buprenorphine) based on the mean AUC_{0-24} values at steady state (Day 28) compared with SL buprenorphine (16 mg once daily for 5 Days) was 31.3%.

Studies in literature of the relationship between plasma buprenorphine levels, opioid receptor occupancy, and the clinical effects of withdrawal suppression and blockade of clinically relevant doses of opioids of abuse feature heterogeneity in the challenge doses used, the interpretation of the term “blockade” (to mean either any detectable attenuation of agonist effect, or complete prevention of agonist effect), and in the doses, route, and timing of the buprenorphine administration. However, although the literature does not support a consistent conclusion about the PK/PD relationships, and the plasma level necessary to provide effective blockade of exogenous opioids at clinically-relevant doses

has not been definitively established, it appears likely to be in the range of 2–3 ng/ml. [8,9] [10,11, 12]

Additionally, the literature confirms that the relationship between buprenorphine plasma levels and the effects of opioid blockade are different from the relationship between buprenorphine levels and withdrawal suppression. A recent article[13] reviewed the scientific data on buprenorphine-induced changes in mu-opioid receptor (μOR) availability, pharmacokinetics and clinical efficacy. The authors concluded that "Opioid withdrawal suppression appears to require ≤50% μOR availability which is associated with BUP plasma concentrations ≥1 ng/ml; for most patients, this may require single daily BUP doses of 4 mg or lower divided doses. Blocking opioid reinforcement requires <20% μOR availability, or BUP plasma levels ≥3 ng/ml for most individuals, this may require single daily BUP doses >16 mg…"

To the extent that blockade of exogenous opioids plays a role in efficacy for addiction treatment, it must be noted that Probuphine is not expected to provide this effect. However, lower plasma levels are required to provide relief of withdrawal symptoms. The exposure provided by Probuphine may be sufficient for this purpose in some patients.

4 Non-Clinical Local Toxicity

Local tissue effects of Probuphine and Placebo (EVA only) implants were evaluated microscopically in dogs after subcutaneous exposures of 1 month and 10 months using standard testing protocols for medical devices.

- Probuphine and Placebo were each moderately irritating after 1 month and slightly irritating after 10 months (see Figure 1).
- Predominant histological observations observed were generally more severe in Probuphine treated animals compared to Placebo animals.
- Severity of local toxicity decreased over time but was substantial during the early phase after implant insertion with the presence of buprenorphine in Probuphine causing increased local toxicity compared to Placebo (see Table 1).

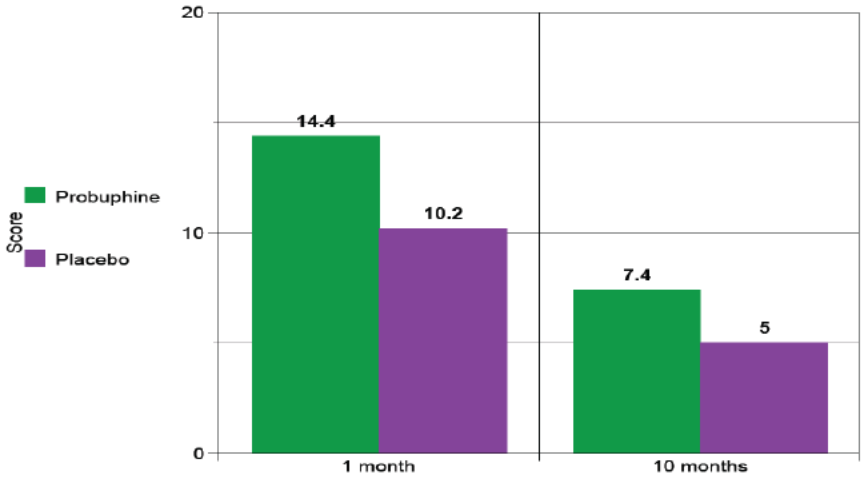
[8] Comer SD, Walker EA, Collins ED. Buprenorphine/naloxone reduces the reinforcing and subjective effects of heroin in heroin-dependent volunteers. Psychopharmacology 2005; 181 (4): 664-675
[9] Strain EC, Walsh SL, Bigelow GE. Blockade of hydromorphone effects by buprenorphine/naloxone and buprenorphine. Psychopharmacology 2002; 159:161-166.
[10] Bickel WK, Stitzer ML, Bigelow GE, Liebson IA, Jasinski DR, Johnson RE. Buprenorphine: dose-related blockade of opioid challenge effects in opioid dependent humans. J Pharmacol Exp Ther. 1988 Oct; 247(1):47-53.
[11] Schuh KJ, Walsh SL, Stitzer ML. Onset, magnitude and duration of opioid blockade produced by buprenorphine and naltrexone in humans. Psychopharmacology (Berl). 1999 Jul; 145(2):162-74.
[12] Mello NK, Mendelson JH, Kuehnle JC. Buprenorphine effects on human heroin self-administration: an operant analysis. J. Pharmacol. Exp. Ther., 223 (1982), pp. 30–39
[13] Greenwald MK, Comer SD, Fiellin DA. Buprenorphine maintenance and mu-receptor availability in the treatment of opioid use disorder: implications for clinical use and policy. Drug And Alcohol Depend. 2014 November 1; 0: 1-11.

Table 1: Severity of Local Toxicity Based on Histological Observations in Dogs Treated with Probuphine or Placebo (EVA only) Implants for 1 or 10 months

Observation	Probuphine		Placebo	
	1 month	10 months	1 month	10 months
Increased fibrosis	1 to 4	1 to 2	1 to 3	1
Increased inflammatory cells				
Polymorphonuclear cells	1 to 3	0	1	0
Lymphocytes	1 to 3	1 to 2	1 to 3	0 to 1
Macrophages	1 to 4	0 to 3	1 to 3	0 to 1
Plasma cells	0 to 1	0 to 1	0	0 to 1

a - severity of fibrosis (0 - none, 1 – narrow band, 2 – moderately thick band, 3 – thick band, 4 – extensive band)
- severity of inflammatory cells (0 – none, 1 – rare, 2 – 5-10 per microscopic field, 3 – heavy infiltrate, 4 – packed)
Table prepared by reviewer from implant site histology data in study PRO-NTR-0215

Figure 1: Individual Implant Local Toxicity Scores



b - Nonirritant (0.0-2.9), Slight Irritant (3.0-8.9)
Moderate Irritant (9.0-15.0), Severe Irritant (≥15.1)
Figure prepared by reviewer from implant site histology data in study PRO-NTR-0215

5 Review of Efficacy

Evidence of efficacy for Probuphine for use in opioid-dependent patients who are clinically stable and on no more than 8 mg a day of sublingual buprenorphine (equivalent to 8 mg/day or less buprenorphine as Suboxone tablet[14]) is derived from a single

[14] For the purposes of this study, equivalent doses include Suboxone, Subutex, or generic tablets. Note that Subutex and Suboxone tablets are no longer marketed by the manufacturer. Equivalent doses of marketed tablets may include generic buprenorphine tablets at a dose of 8 mg buprenorphine; generic buprenorphine/naloxone tablets 8 mg/2 mg buprenorphine/naloxone; Zubsolv tablets at a dose of 5.7

controlled trial, PRO-814. PRO-814 was a randomized, double-blind, double-dummy, active-controlled, multicenter trial that evaluated the safety and efficacy of four 80-mg Probuphine implants in adult outpatients with opioid dependence who were on ≤8 mg SL buprenorphine and considered clinically stable by their treating healthcare provider.

5.1 Study Design and Endpoints

The study was a randomized, double-blind, double-dummy, active-controlled trial. Eligible participants included patients 18 to 65 years of age who met DSM-IV criteria for opioid dependence as a primary diagnosis and were considered clinically stable by their treating healthcare provider. Clinical stability was confirmed by the following criteria at the time of randomization: (1) on sublingual (SL) buprenorphine treatment for at least 6 months (≥ 24 weeks); (2) on a stable SL buprenorphine dose ≤ 8 mg daily for at least the last 90 days; and (3) no positive urine toxicology for illicit opioids in last 90 days. Eligible patients were also to be free from significant withdrawal symptoms (COWS score ≤ 5). To document clinical stability, the treating healthcare providers were to complete a Clinical Stability Checklist (see Appendix C).

The definition of "clinically stable" for the opioid-dependent population (DSM-IV criteria) is not a well-established and there are not criteria that are universally understood to define clinical stability. During communications with the Applicant, the Agency stressed that in defining a population for the study, it was important to note that being on a *stable dose* of SL buprenorphine is not synonymous with being *clinically stable*. In defining the population and responder definition for the trial, the Applicant reviewed existing literature on patients whose clinical picture appeared consistent with clinical stability on buprenorphine treatment. Additionally, the Applicant conducted a survey of addiction experts which included the following questions.

1. In this same patient population, how often do you expect the average stable patient in your practice to test positive for opioids over a 6-month period?

2. If these patients were to continue on the same dose, what would be the overall average percentage of opioid-negative urine toxicology result would you anticipate in 6 months?

3. If their buprenorphine treatment were to be stopped, what would be the average percentage of relapse in these patients over a 6-month period?

4. Assume urine toxicology is measured monthly for six months. What would you consider to be the maximum reasonable change in a stable patient's urine toxicology status, for the patient to continue to be considered stable?
   a. No change
   b. 1 out of 6 urine-positive urine toxicology
   c. 2 out of 6 urine-positive urine toxicology

---

mg/0.71 buprenorphine/naloxone; Bunavail buccal film at a dose of 4.2 mg/0.7 mg buprenorphine/naloxone. Suboxone Film, 8 mg/2 mg buprenorphine/naloxone delivers a somewhat higher exposure to buprenorphine than Suboxone tablets at the same dose.

d. 3 or more out of 6 urine-positive urine toxicology

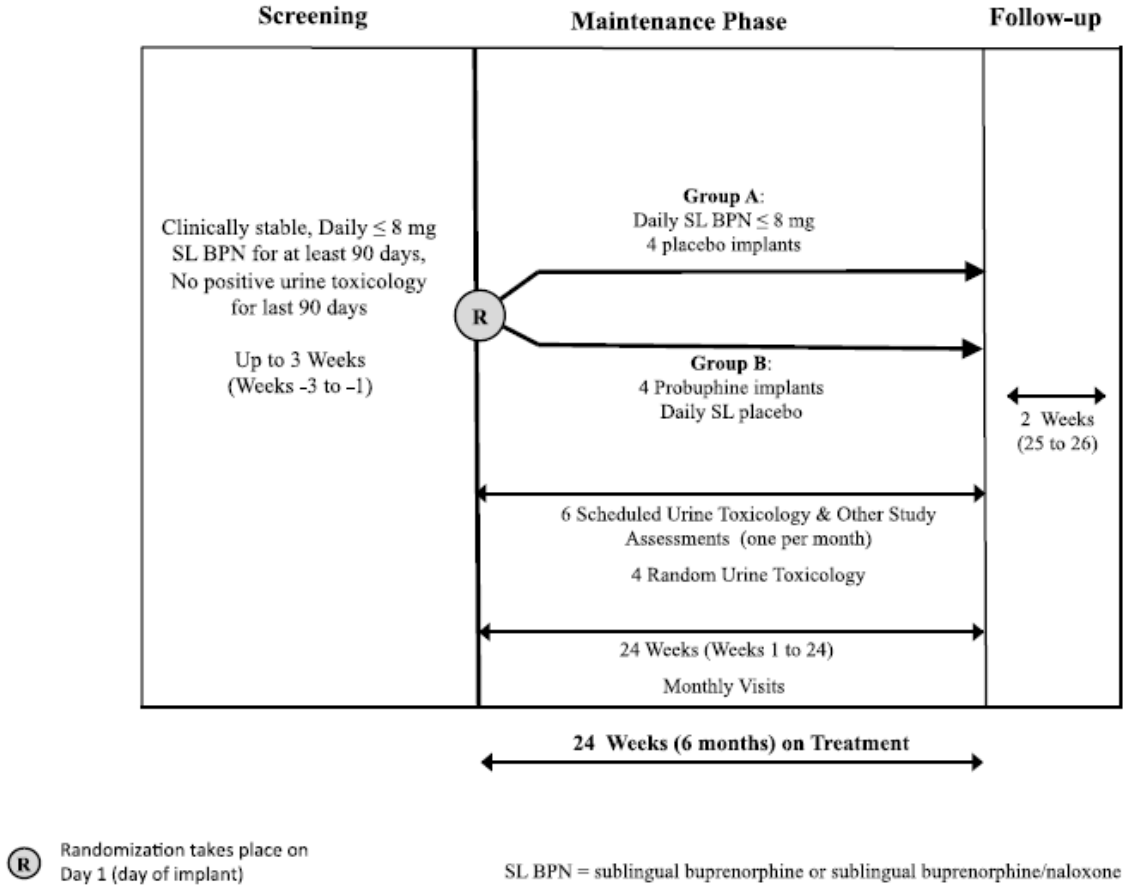
5. What percentage of stable patients do you expect would fall into the above category? (i.e., to have the maximum reasonable change in urine toxicology status over six months)

Both the literature and physician survey data explored by the Applicant in formulating a responder definition described patients who were on buprenorphine treatment for at least six consecutive months. However, although the protocol for PRO-814 stipulated that patients needed to have been on "sublingual buprenorphine treatment for at least 6 months (≥ 24 weeks)," the Applicant did not include measures in the protocol to ensure that only patients with at least six consecutive months of buprenorphine treatment would be enrolled, and the criterion was apparently interpreted by investigators to mean cumulative lifetime duration of buprenorphine treatment. Therefore, the study included both patients in long-term, stable treatment (as intended), and some with treatment episodes of shorter duration prior to entry. Additionally, the protocol stipulated that patients were to be on a stable SL buprenorphine dose ≤ 8 mg daily for at least the last 90 days. However, it should be noted that there are a number of transmucosal-buprenorphine containing products, which are not all bioequivalent, so defining patients by baseline dose introduces some uncertainty.

Patients were ineligible if they required opioid treatment of a current chronic pain condition, met criteria for dependence on other psychoactive substances (nicotine dependence permitted), or used illicit benzodiazepines. Medical reasons for exclusion included elevated hepatic enzymes, bilirubin, or creatinine; low platelets; coagulopathy or anticoagulant treatment; recent scarring or tattoos on upper arm, or history of keloid scarring; use of CYP3A4 inhibitors; a current AIDS diagnosis; or other medical or psychiatric conditions at investigator discretion.

Eligible patients were to participate in three phases over the course of the trial: a Screening Phase (up to 3 weeks in duration), a 24-week Maintenance Phase; and a 2-week Follow-Up Phase. The following figure provides an overview of the study design for the trial.

Figure 2: Overview of Study Design



On Day 1 of the trial, subjects were to be randomized in a 1:1 ratio to 1 of 2 treatment groups:

> **Group A**: Sublingual Buprenorphine tablets (≤ 8 mg/daily) + four placebo implants
> **Group B**: Four 80-mg Probuphine implants + daily SL placebo tablets

On Day 1, Probuphine and placebo implants were to be implanted by a trained Implant Clinician at the clinical study site.[15] Prior to randomization on Day 1, it was recommended that subjects discontinue their prior sublingual buprenorphine and have implants inserted subdermally within 12–24 hours after their last sublingual buprenorphine dose. Insertion under the skin of the upper arm was performed using a specialized applicator provided by the Applicant, (which the Applicant describes as similar in design to commercially-approved applicators used for the insertion of other implantable drugs). Subjects were to be monitored closely for AEs and vital signs for at least 30 minutes following insertion by medically qualified personnel.

[15] The clinicians who inserted the Probuphine rods across the trial sites included 8 Family Medicine, 6 Obstetrics & Gynecology, 3 General Surgery, 3 Anesthesiology, 1 Neurology and Psychiatry, and 1 Radiation Oncology physicians, and 1 Family Practice Nurse Practitioner.

Subjects randomized to sublingual buprenorphine tablets plus placebo implants were transitioned to the same dose of sublingual buprenorphine on which they were previously maintained, using a generic version of buprenorphine/sublingual tablets bioequivalent to Suboxone tablets.

After Day 1, subjects were to return to the site for 7 scheduled Maintenance Phase visits. Maintenance Phase visits were scheduled to occur at Week 1 (Post-Implant Follow-Up Visit) and then monthly from Week 4 through Week 24 (End of Treatment [EOT] Visit). Subjects also were to be scheduled to visit the site 4 other times during the Maintenance Phase to provide random urine toxicology samples. The Post-Treatment Telephone Contact (Week 25) and the Follow-Up Visit (Week 26) occurred 1 week and 2 weeks after the EOT Visit, respectively. Implants were removed at the EOT Visit.

At each maintenance visit, the following procedures and assessments were to be performed: psychosocial counseling, urine toxicology, illicit drug use self-report, withdrawal assessments (SOWS, COWS), Desire to Use/Need to Use VAS, dispense sublingual buprenorphine or placebo (except EOT visit), AE assessment, vital signs, pregnancy test, implant site examination, concomitant medication assessment.

During the Maintenance Phase, a total of 10 urine toxicology samples were to be collected, 6 at regularly scheduled monthly visits and 4 at randomly scheduled visits. It was recommended that the urine toxicology samples be collected on Mondays to potentially improve detection of illicit opioid use that may have occurred over the weekend. In scheduling random urine toxicology visits, sites were advised to utilize their standard clinical process for scheduling the 4 random urine drug testing visits. It was recommended, however, that no more than one random urine test be conducted between two scheduled visits. Each of the random visits was to occur within 48 hours after speaking directly with the subject.

Urine samples were to be logged and numbered and then sent to the central laboratory for quantitative analysis (via LCMS/ MS methods) of opioids (codeine, morphine, hydrocodone, oxymorphone, hydromorphone, oxycodone, methadone, dihydrocodeine, and fentanyl) and opioid metabolites (EDDP [methadone metabolite], and norfentanyl). Investigators were privy to the urine toxicology test results during the trial, and generally received the results 7 to 10 days from the date the sample was received by the central laboratory.

The protocol permitted provision of supplemental SL buprenorphine and other interventions. Investigators were instructed to treat additional symptoms of opioid dependence as they normally would, including additional counseling sessions, supplemental SL BPN, or other pharmacological interventions. It is important to note that subjects were told that while additional counseling and other pharmacological interventions were available, their then-current dose of BPN was expected to be adequate to maintain stability and that they were not expected to need supplemental SL BPN. In this trial, it was anticipated that supplemental sublingual buprenorphine use was to be so

infrequent and so sporadic so as not to warrant protocol-specified criteria for administration, or factoring supplemental use into treatment responder or failure definitions.

Any supplemental SL buprenorphine, additional counseling, and other pharmacological interventions provided by the investigator were to be recorded, along with the reasons for determining the need for supplemental intervention.

Protocol-specified criteria for early withdrawal from the study included the following: subject request; subject non-compliance (defined as refusal or inability to adhere to the study protocol); evidence of implant removal or attempted implant removal; pregnancy; intercurrent illness that, in the judgment of the investigator, affected assessments of clinical status to a significant extent, required discontinuation of drug, or both; requirement for continual use of opioid analgesics >7 days or general anesthesia for surgery; at the request of the Applicant, regulatory agencies, or the Institutional Review Board (IRB); or loss to follow-up.

The primary efficacy endpoint for this study was the proportion of responders for the Probuphine and sublingual buprenorphine treatment groups. The definition of responder for this study was any patient with no more than 2 out of 6 months with any evidence of illicit opioid use. Evidence of illicit opioid use was defined as either a positive opioid urine toxicology result or self-reported illicit opioid use.

During discussions with the Applicant about the NDA resubmission, there was consensus that the primary efficacy endpoint would be the proportion of subjects in each treatment group with no or minimal change from baseline based on the percentage of opioid-negative urine toxicology results. The Applicant initially proposed to operationalize the responder definition as subjects with percentage of opioid-negative urine tests adjusted for self-reported drug use over 24 weeks (12 scheduled samples with more frequent testing in a baseline phase comprising the first month and consisting of weekly visits, and during the last four weeks of the 24-week ascertainment window that followed the baseline, as originally proposed) demonstrating no or minimal change (up to 16.7% as per addiction specialist survey) from baseline. As the weekly testing during the Baseline period and the last four weeks of the 24-week ascertainment window appeared inconsistent with clinical management for stable patients and the absolute threshold for deterioration was too permissive, the Applicant was advised that the responder definition should be revised. On the Division's suggestion for an alternate approach, the Applicant agreed to conduct scheduled monthly urine toxicology visits with interspersed random urine visits, which was considered to align more appropriately with clinical management for clinically stable patients. The responder definition would then be any patient with no more than 2 out of 6 months with any evidence of illicit opioid use, with evidence of illicit opioid use defined as above.[16]

[16] In discussing the responder definition with the sponsor, the Agency provided the following comment: The responder definition should be revised. As currently proposed, the definition is too permissive in that it sets the absolute threshold for deterioration to allow a patient to have as few as 58.3% samples negative over a six-month period of time, in patients who are deemed clinically stable by their

healthcare provider. In this definition, clinical stability as evidenced by submitting no more than one urine sample positive for opiates over a six-month period of monthly testing is inappropriately converted to a percentage of negative urines collected over a six month period at a more frequent schedule.

The definition also includes an allowance for change from the observed baseline. However, the observed baseline is only four weeks long and requires patients to attend weekly visits. Among clinically-stable patients, opioid-positive urine toxicology findings are anticipated to be infrequent, and to occur sporadically over a longer timeframe than the four weeks proposed for the Baseline Phase, e.g., six months or longer. As such, assessments made during the Baseline Phase appear to underestimate clinical stability because the four-week period is not an adequate amount of time with which to essentially reassess clinical stability in patients enrolled based on a status of clinical stability. To avoid underestimating clinical stability during the Baseline Phase, missing visits and, in turn, missed urine sample collections should be adjudicated as negative in these clinically-stable patients, and the randomization criteria to be met at the end of this phase should be modified such that only those with 100% opioid-negative urine toxicology results should be eligible for randomization.

Recommendations on enhancing the Baseline Phase aside, the purpose of the proposed Baseline Phase, which adds an additional four visits to the trial, is not obvious, and we ask that you provide your rationale for including this phase.

Because stable patients are usually assessed once monthly for six samples over a six-month period, as opposed to the 12 samples proposed over this same period, we recommend that, to more closely approximate the clinical assessments, the responder definition should be defined by month, as opposed to percentage of opioid-negative urine samples collected over the observation period. Similarly, the maximum reasonable change endorsed by surveyed addiction specialists was 1/6 samples and assumed monthly urine toxicology for six months, that is 6 samples across six months vs. the 12 samples proposed for the study. Accordingly, we recommend that the responder definition should be based on opioid-free months, where an opioid-free month is one in which there is no evidence of illicit opioid use either by urine toxicology or self-report. Taking into account the expectation that clinically-stable patients will have on average no more than one opioid-positive urine sample in a six month period, and the clinicians' perspective on the maximum reasonable change (one additional month with a positive sample), a responder should be defined as a patient with no more than two months with any evidence of illicit opioid use.

5.2 Population

A total of 177 subjects were enrolled and randomized into the study at a total of 21 sites within the United States. Of these, 176 subjects received study medication and were included in the safety population (89, SL BPN; 87, Probuphine).

Selected demographic and baseline characteristics of the patients are shown in the tables below.

Table 2: Demographics

Category	SL BPN N=89	Probuphine N=87	Total N=176
Age (years)			
Mean (SD)	39 (10.8)	38 (11.2)	39 (11.0)
Min, max	22, 64.0	21, 63.0	21, 64.0
Gender, n (%)			
Male	52 (58.4)	52 (59.8)	104 (59.1)
Female	37 (41.6)	35 (40.2)	72 (40.9)
Race, n (%)			
White	85 (95.5)	82 (94.3)	167 (94.9)
Black or African American	2 (2.2)	3 (3.4)	5 (2.8)
Asian	0	1 (1.1)	1 (0.6)
American Indian or Alaska native	1 (1.1)	0	1 (0.6)
Other	1 (1.1)	0	1 (0.6)
Ethnicity, n (%)			
Hispanic or Latino	3 (3.4)	3 (3.4)	6 (3.4)
Not Hispanic or Latino	86 (96.6)	84 (96.6)	170 (96.6)
BMI (kg/m2)			
Mean (SD)	27 (5.92)	28 (6.94)	28 (6.47)
Min, max	19, 50.6	14, 46.4	14, 50.6

Abbreviations: BMI, body mass index; SD, standard deviation; SL BPN, sublingual buprenorphine
Source: PRO-814 CSR, Table 7, p. 57

Table 3: History of Opioid Abuse

Category	SL BPN N=89	Probuphine N=87
Met DSM-IV-TR criteria for opioid abuse, n (%)		
Yes	89 (100.0)	86 (98.9)
Not reported[a]	0	1 (1.1)
Primary opioid of abuse, n (%)		
Prescription opioid pain reliever	65 (73.0)	66 (75.9)
Heroin	22 (24.7)	15 (17.2)
Other	2 (2.2)	5 (5.7)
Not reported	0	1 (1.1)
Time since first opioid abuse (years)		
N	89	86
Mean (SD)	11.5 (7.68)	11.2 (6.62)
Median	10.7	10.1
Min, max	1.6, 45.6	1.4, 36.6
Time since first diagnosis (years)		
N	89	86
Mean (SD)	6.2 (6.95)	6.2 (5.93)
Median	3.9	5.4
Min, max	0.6, 43.6	0.5, 34.6

Abbreviations: DSM-IV, Diagnostic and Statistical Manual – 4th Edition – Text Revision; SL BPN, sublingual buprenorphine

[a] For Subject 012-006 (Probuphine), the substance abuse and substance abuse treatment histories were not completed at the screening visit

Table 4: Subject Disposition

Category	SL BPN	Probuphine	Total
Randomized	90	87	177
Safety Population (N)	89	87	176
Completed, n (%)	84 (94.4)	81 (93.1)	165 (93.8)
Discontinued, n (%)	5 (5.6)	6 (6.9)	11 (6.3)
Reason for discontinuation, n (%)			
Adverse event	0	1 (1.1)	1 (0.6)
Request of sponsor or regulatory agency	1 (1.1)	0	1 (0.6)
Lost to follow-up	2 (2.2)	4 (4.6)	6 (3.4)
Other[a]	0	1 (1.1)	1 (0.6)
Subject request	2 (2.2)	0	2 (1.1)

Abbreviations: SL BPN, sublingual buprenorphine
[a] Subject was incarcerated and not able to complete the study visits.

Table 5: Study Populations

Category	SL BPN	Probuphine	Total
Randomized (N)	90	87	177
Safety population, n (%)	89 (98.9)	87 (100.0)	176 (99.4)
ITT population, n (%)	89 (98.9)	84 (96.6)	173 (97.7)
PP population, n (%)	72 (80.0)	67 (77.0)	139 (78.5)

Abbreviations: ITT, intent-to-treat; PP, per-protocol; SL BPN, sublingual buprenorphine

5.3 Statistical Methodologies

5.3.1 Historical Effect of Sublingual Buprenorphine and Choice of Non-Inferiority Margin

This study evaluated the efficacy of Probuphine in opioid dependent patients that were considered clinically stable of 8 mg or less of buprenorphine. The Applicant and the Division agreed it would be inappropriate to put these stable patients at risk of relapse by conducting a placebo-controlled trial. Therefore, the Division agreed to a non-inferiority (NI) design.

According to the draft Guidance issued by the Agency in 2010, this study should show that the difference, with respect to efficacy, between Probuphine and sublingual buprenorphine is small enough to allow the known efficacy of sublingual buprenorphine to support the conclusion that Probuphine is also effective. The smallest decrease in effect from that of sublingual buprenorphine that would be acceptable is referred to as the Non-Inferiority (NI) margin. If the lower bound of the confidence interval for the effect of Probuphine compared to sublingual buprenorphine is greater than this margin, then Probuphine would be considered non-inferior to sublingual buprenorphine. There are two important questions to answer to determine the non-inferiority margin, what is the historical response rate for sublingual buprenorphine and how small a difference between Probuphine and sublingual buprenorphine would be acceptable?

To provide evidence of the historical effect of sublingual buprenorphine, the Applicant provided four publications and results from a survey of addiction specialists to provide an estimates of the historical response rate for sublingual buprenorphine following blinded taper or complete withdrawal from long term buprenorphine or methadone treatment. The references were summarized by the Applicant as follows:

> A meta-analysis of tapered discontinuation following long-term methadone or BPN treatment found an average abstinence rate of 33% (Korner & Waal, 2005)[17]. However, because of the differences in methodology (single or double-blinding, naturalistic, etc.), definitions of abstinence, treatments administered during MAT and durations of follow-up, some studies are more relevant than others. In addition, this article didn't report on the baseline rates of percentage abstinence or urine toxicology results.
>
> Breen et al., (2003)[18] reported on a study of stable methadone patients (for at least 6 months) to BPN and then gradual reduction to 0 mg BPN (i.e., blinded) over an average duration of 11 weeks showed that subjects at 1

[17] Kornør H, Waal H. From opioid maintenance to abstinence: a literature review. Drug Alcohol Rev. 2005 May;24(3):267-74.

[18] Breen CL, Harris SJ, Lintzeris N, Mattick RP, Hawken L, Bell J, Ritter AJ, Lenne M, Mendoza E. Cessation of methadone maintenance treatment using buprenorphine: transfer from methadone to buprenorphine and subsequent buprenorphine reductions. Drug Alcohol Depend. 2003 Jul 20;71(1):49-55.

month follow-up after complete BPN discontinuation had 31% negative opioid samples (relative to about 73% negative at baseline, 89% negative during BPN induction, and 91% negative during BPN taper).

One double-blind, double-dummy study in methadone users found 25% abstinence overall during 1 month follow-up after complete discontinuation following gradual taper regimens. Abstinence was 18% in the "rapid" withdrawal group (taper over 10 weeks) (versus 100% negative urine opioid results for 4 weeks preceding study entry and 92% negative for the 6 months prior to the study) (Senay, 1977)[19].

Most of the studies used tapered discontinuation, but in terms of abrupt discontinuation, one survey study in Australia reported that 15% of patients who abruptly discontinued opioid maintenance therapy (BPN or methadone) were abstinent for at least 3 months, while 26-27% were abstinent with either self- or physician-directed taper regimens (Winnstock et al., 2011)[20].

The results of the survey of addiction specialists are summarized in Table 6: Summary of Survey Results from Addiction Specialists. The Applicant found that the addictions specialists estimated that a median of 25% of subjects who were stabilized on 8 mg or less of sublingual buprenorphine would not relapse upon discontinuation of their buprenorphine dose.

[19] Senay EC, Dorus W, Goldberg F, Thornton W. Withdrawal from methadone maintenance. Rate of withdrawal and expectation. Arch Gen Psychiatry. 1977 Mar;34(3):361-7.

[20] Winstock AR, Lintzeris N, Lea T. "Should I stay or should I go?" Coming off methadone and buprenorphine treatment. Int J Drug Policy. 2011 Jan;22(1):77-81. doi: 10.1016/j.drugpo.2010.08.001. Epub 2010 Oct 16.

Table 6: Summary of Survey Results from Addiction Specialists

PI #	% Negative UDS Over 6 mths*	% Negative UDS Over Next 6 mths*	% Relapse upon BPN Discontinuation (Over 6 mths)	Maximum Reasonable Change in % UDS Positive	% of Patients
1	100	83	85	0	85
2	75	80	75	17	65
3	NNR	DNAQ	NNR	33	55
4	NNR	90	35	17	10
5	75	91.5	70	17	60
6	NNR	NNR	NNR	17	55
7	99	95	90	17	95
8	80	95	80	17	20
9	95	95	80	8.5	80
10	NNR	65	75	33	60
11	83	90	75	17	66
12	100	80	95	0	85
13	100	100	60	0	10
14	100	90	30	0	90
15	91.5	90	75	17	60
16	91.5	90	50	17	80
17	100	90	90	17	90
18	100	100	60	8.5	NNR
Mean (Median)	92 (97)	89 (90)	70 (75)	14 (17)	63 (65)
Range	75-100	65-100	30-95	0-33	10-95

NOTES: DNAQ =response given did not match question asked and is not useful for the averages;
NNR =no numerical response; UDS=Urine Opioid toxicology

* Some answered as % positive some as % negative, for ease, results have been converted to % negative.

- If range was given; the average of the range has been entered here (i.e., 30-40% = 35% for purposes of these calculations)
- If answer given as < or > , response was entered as the numeric value
- "X of 6 responses were calculated as: 0 of 6 = 0%; 1 of 6 = 17%; 2 of 6 = 33%; 3 of 6 = 50%; 4 of 6 = 67%; 5 of 6 = 83%; 6 of 6 = 100%

The Applicant estimated that the response rate for subjects receiving placebo would be 25% and the effect size for patients who continued buprenorphine treatment would be 75%. The effect size is the difference between the comparator (placebo) response rate and the treatment group response rate; therefore this effect size implies an assumption that all subjects who continue to receive their sublingual buprenorphine dose would remain stable, i.e. the response rate would be 100%. Using this estimate, the Applicant selected a NI margin of 20% for this study. This margin was chosen to preserve at least 70% of their estimate effect size of the sublingual buprenorphine which they considered to be clinically acceptable.

5.3.2 Primary Analysis

The Applicant stated that the primary analysis population was the Intent-to-Treat (ITT) Population; however, there were two different definitions for this population. The study protocol defines this population as "All subjects who have been randomized and have received an implant and/or received sublingual buprenorphine/placebo". The statistical analysis plan and the final study report both use the same definition for this population which is, "All randomized subjects who received study medication and provided some efficacy data". We consider the first definition to be appropriate.

The Applicant's primary efficacy endpoint for this study was the proportion of responders for the Probuphine and sublingual buprenorphine treatment groups. The Applicant defined a responder as a patient with no more than 2 out of 6 months with any evidence of illicit opioid use. However, the suitability of this definition was questionable given the quantity of supplemental medication use observed in this study (see Section 5.4.3). It was therefore necessary to explore different definitions of responder that incorporate the level of supplemental medication in their definition.

To determine if Probuphine was NI to sublingual buprenorphine the following procedure was followed: Let π_c and π_t represent the proportion of responders at 24-weeks for the active control arm (sublingual buprenorphine) and the experimental treatment arm (Probuphine), respectively. The null hypothesis of inferiority can be stated as:
H_0: $\pi_t \leq \pi_c - 0.20$
The alternative hypothesis of non-inferiority can then be stated as:
H_A: $\pi_t > \pi_c - 0.20$
This hypothesis can then be tested by checking whether the lower bound of the 95% confidence interval for the difference between Probuphine and sublingual buprenorphine is larger than -0.20. If this is true then the null hypothesis of inferiority can be rejected and non-inferiority can be concluded. The Applicant used the standard Wald confidence intervals for this test.

5.3.3 Handling of Missing Data

The primary imputation utilized by the Applicant involved a 20% (non-inferiority margin) relative penalty. With this imputation scheme, missing urine values in the sublingual buprenorphine group were imputed based on the proportion of "opioid-positive" samples within the treatment group. This proportion was the average of the within-subject proportion of "opioid-positive" sample. Missing urine values in the Probuphine group were imputed based on the proportion equal to 1.2 times the maximum proportion of the 2 proportions from the 2 treatment groups. For example, if the proportions of "opioid-positive" samples were to be 14% for the Probuphine group and 15% for sublingual buprenorphine group, the imputation for the sublingual buprenorphine group would be based on 15% and the imputation for the Probuphine group would be based on 18%.

The primary underlying assumption for this analysis is that the likelihood of missingness is unrelated to the subject's illicit opioid usage. The Applicant provided no justification or exploration of the plausibility of this assumption and did not conduct any sensitivity analyses to explore the effect of varying this assumption on the conclusion of the study. In order to explore the extent of the effect of missing data on the conclusion it was necessary to conduct additional sensitivity analyses that explore a range of plausible assumptions on the missing data structure. The following sensitivity analyses were conducted.

- All missing urine toxicology tests were imputed to be positive. While the Applicant believes that imputing the missing observations using extreme values

will introduce bias it is necessary to explore such scenarios from a regulatory viewpoint as they represent plausible outcomes in this study and population and we cannot be confident in the conclusions of this study unless we explore such scenarios.

- Any urine toxicology test that was considered inconclusive was imputed as positive. There were a number of urine toxicology tests in the study where the results of the individual tests within the opioid panels were inconclusive due to a variety of reasons which will be discussed further in Section 5.4.2.

5.4 Results and Conclusions

All results presented utilized the ITT population defined as all randomized and treated patients. Subjects that discontinued prior to Week 24 were considered non-responders. There were three concerns noted during the review of this submission that resulted in further analyses by the Agency; the analysis population, number of subjects with missing and inconclusive urine samples, and use of supplemental sublingual buprenorphine. These are discussed separately in Sections 5.4.1, 5.4.2, and 5.4.3, respectively.

An investigation of the success rates by prior dose is presented in Section 5.4.4 and a discussion of the non-inferiority margin is presented in Section 5.4.5.

5.4.1 Influence of Analysis Population

The Applicant's primary analysis excluded three subjects in the Probuphine arm that received study drug but discontinued without providing any efficacy data. Exclusion of these subjects from the analysis population resulted in the Applicant concluding that Probuphine was both non-inferior and superior to sublingual buprenorphine. When these three subjects were included in the analysis as non-responders, although NI was still evident, superiority was no longer established. The 95% Wald confidence intervals for the difference in the proportion of responders for Probuphine and sublingual buprenorphine and the p-value for the test of superiority are shown in the table.

Table 7: Influence of Analysis Population

Category	Probuphine n (%)	SL BPN n (%)	Proportion Difference (95% CI) Probuphine – SL BPN	*P* Value (2-Sided)
Applicant's Primary Analysis				
N	84	89		
Responder	81 (96%)	78 (88%)	0.088 (0.009, 0.167)	0.03
Non-responder	3 (4%)	11 (12%)		
Applicant's Primary Analysis Using Protocol Definition of ITT Population				
N	87	89		
Responder	81 (93)	78 (88)	0.055 (-0.032, 0.141)	0.22
Non-responder	6 (7)	11 (12)		

5.4.2 Missing and Incomplete Urine Toxicology Results

The Applicant planned a total of ten urine toxicology tests per subject enrolled in the study in the double-blind treatment phase, six during the monthly clinic visits and four random tests. Each of the urine toxicology panels consisted of a total of 17 tests for illicit substances, of which 11 were for either opioids or opioid metabolites, and 6 were for other non-opioid illicit substances. The Applicant also reported the creatinine concentration and opioid/opioid metabolite creatinine ratios. If any quantity of an opioid or opioid metabolite was detected then that test was classified as positive. The results of the non-opioid illicit substance tests were not utilized in the primary analysis.
The overall results of the urine toxicology tests are shown in Table 8. There were two issues that caused missing data in this trial. As would be expected there were a number of missed visits, the majority of which were for subjects who discontinued from the study before completion. There were also a number of visits where a sample was collected, but due to various issues it was not possible to obtain results for all the planned opioid tests
.

1. Norfentanyl content was unable to be determined due to "matrix problems" (66 visits).
2. Creatinine concentration and opioid/creatinine ratios were unable to be determined (17 visits).
3. Creatinine and all opioids except methadone or fentanyl were out of stability and unable to be analyzed (15 visits).
4. Oxymorphone results were not reported due to interference by a morphine metabolite (8 visits).
5. Subjects were missing non-opioid tests (6 visits).

If at least one result was reported as positive then that test was reported as positive regardless of any missing results. However, the majority of tests with missing results were reported as negative as no opioids were detected.

Table 8: Urine Toxicology Results by Treatment Group

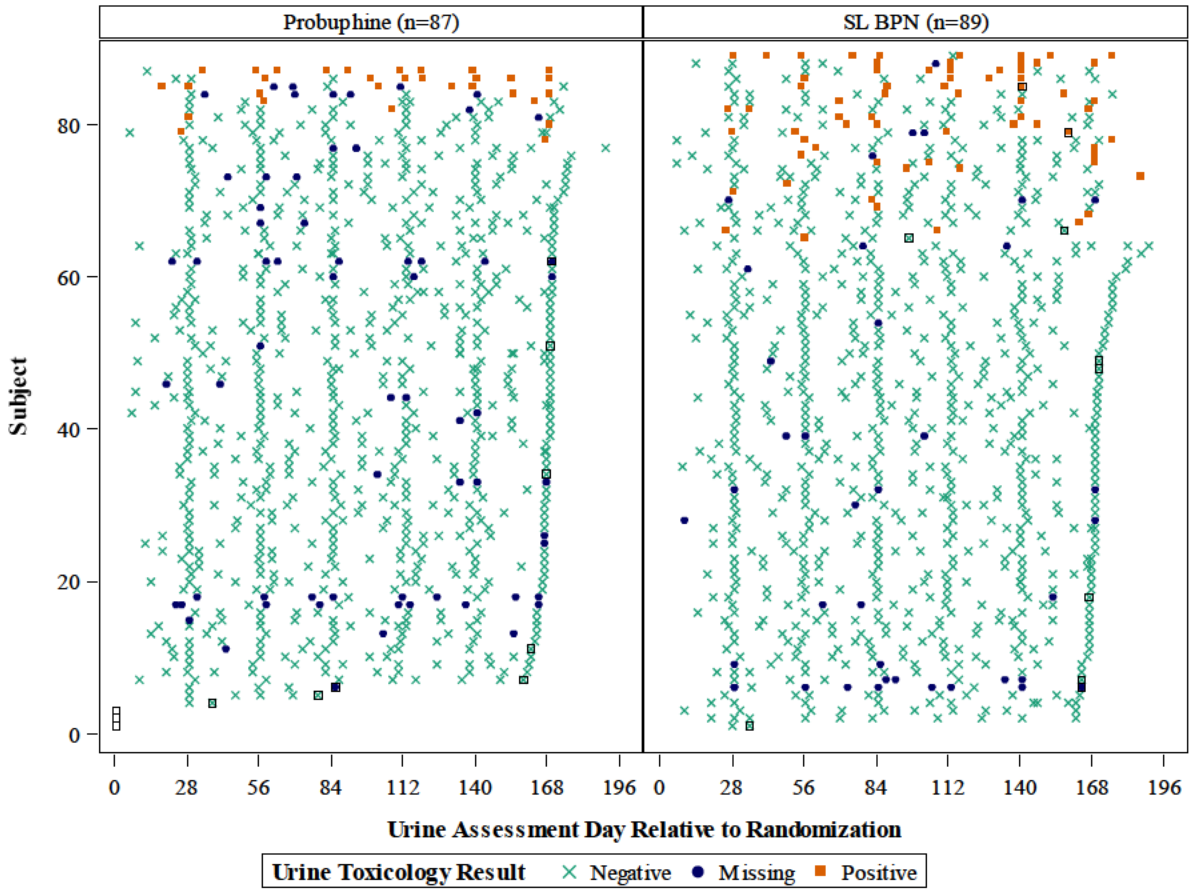
Treatment Group	Negative n (%)	Positive n (%)	Incomplete Result n (%)	Missing Sample n (%)	Total
SL BPN (n=89)	765 (86%)	64 (7%)	34 (4%)	27 (3%)	890
Probuphine (n=87)	725 (83%)	31 (4%)	60 (7%)	54 (6%)	870

SL BPN = sublingual buprenorphine

Figure 3 shows the distribution of urine toxicology results by subject. The blue circle indicates visits where there were either missing individual test results or a sample was not provided. However, the majority of missing samples are omitted from this plot since they were scheduled after subjects discontinued from the study. We also see from Figure 3 that the missing individual test results and missing samples are concentrated in certain subjects rather than evenly distributed across all the subjects in the study. This appears to

be due to the matrix problems with the norfentanyl analysis which occur repeatedly for the same subjects[21].

Figure 3: Urine Toxicology Test Results



Note: Black squares mark the final visit for subjects who did not provide all 10 urine toxicology samples.
SL BPN = sublingual buprenorphine

Table 9 summarizes the number and proportion of subjects in the study with the specified issues.

Table 9: Number (%) of Subjects with Specified Issue

Issue	Probuphine n (%)	SL BPN n (%)	Total n (%)

[21] The "matrix" for these tests is the urine sample. The Applicant explained that "matrix problems can occur when there are certain samples that may contain other compounds such as natural breakdown products within the urine or potential concomitant medications that the subjects were taking that interfere with chromatography of the lab's methods. Therefore, the test process was affected and definitive results could not be reported." It is not clear whether these problems can also be a result of deliberate tampering or attempts to avoid detection.

N	87	89	176
No Issues	46 (53%)	49 (55%)	95 (54%)
No Missing Data	56 (64%)	67 (75%)	123 (70%)
Missed Sample	11 (13%)	11 (12%)	22 (13%)
Incomplete Result	22 (25%)	16 (18%)	38 (22%)
Rescue Use	15 (17%)	13 (15%)	28 (16%)
Positive Test	10 (12%)	25 (28%)	35 (20%)

SL BPN = sublingual buprenorphine

Additional analyses were conducted to examine the impact of missing urines samples and inconclusive urine toxicology tests. Table 10 shows the percentage of responders in each treatment arm and the corresponding CI for the difference when missing urine sample are considered positive and when missing urine samples and incomplete urine test results are considered positive. The p-value for the test of superiority is also included.

Table 10: Missing Data Analysis

Category	Probuphine n (%)	SL BPN n (%)	Proportion Difference (95% CI) Probuphine – SL BPN	*P* Value (2-Sided)
Missing Urine Samples Imputed as Positive				
N	87	89		
Responder	78 (90%)	76 (85%)	0.043 (-0.055, 0.140)	0.39
Non-responder	9 (10%)	13 (15%)		
Missing Urine Panels Imputed as Positive				
N	87	89		
Responder	73 (84%)	70 (79%)	0.053 (-0.062, 0.167)	0.37
Non-responder	14 (16%)	19 (21%)		

SL BPN = sublingual buprenorphine

In both cases NI of Probuphine to sublingual buprenorphine was established. There was no evidence of superiority.

5.4.3 Use of Supplemental Sublingual Buprenorphine

Supplemental buprenorphine was dispensed as 2 mg sublingual tablets. The Applicant did not record the dates when the supplemental medication was actually used, only when and how much was dispensed. The Applicant reported the dose for only a single subject who received an additional 4 mg per day. It was assumed that all other subjects used a single additional 2 mg tablet per day.

The number of subjects who required supplemental sublingual buprenorphine, the total number of dispensing episodes, and the average number of tablets dispensed per subject receiving sublingual buprenorphine are shown in Table 11. The percentage of subjects and the total number of dispensing episodes is relatively similar between the two arms; however, the average number of tablets per subject was much larger in the Probuphine treatment arm than the sublingual buprenorphine treatment arm.

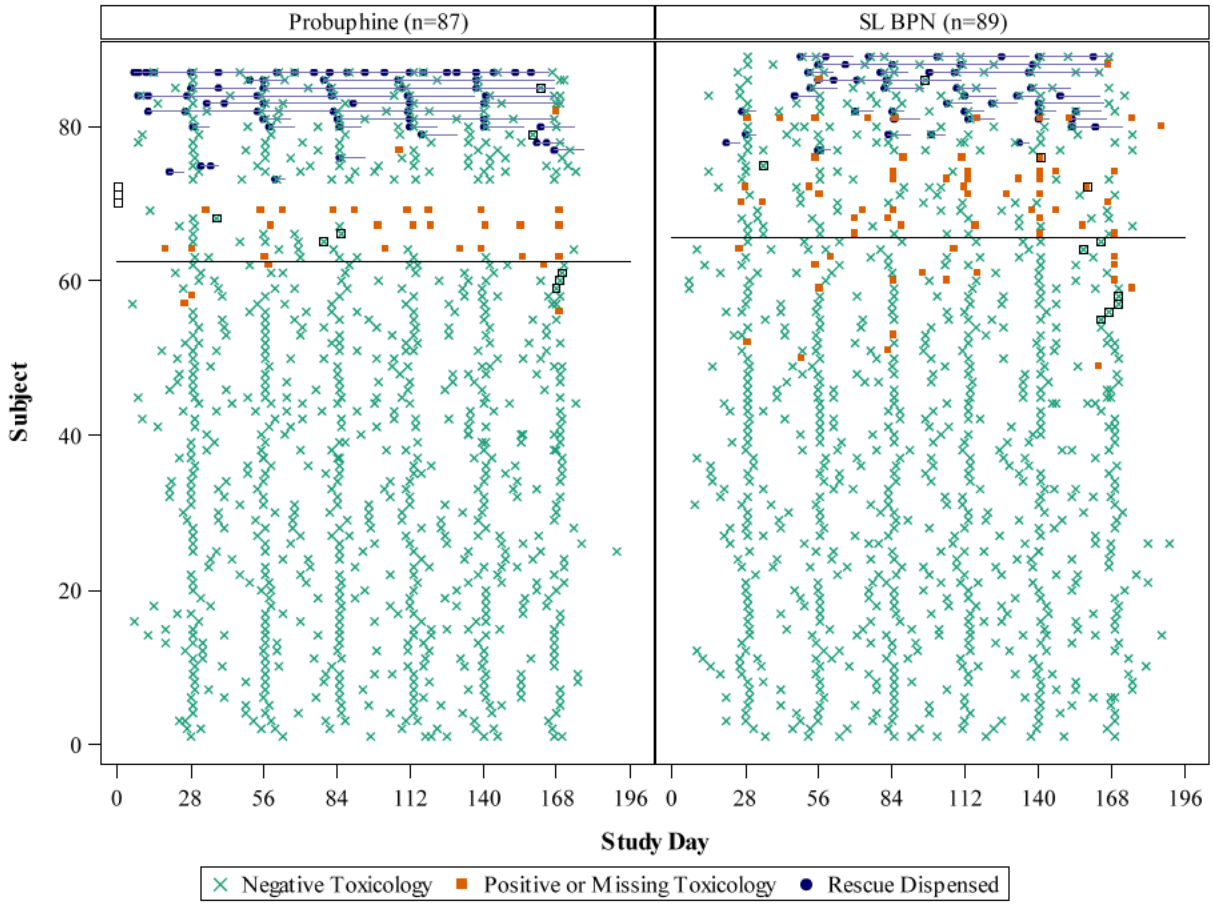
Table 11: Summary of Supplemental Sublingual Buprenorphine Usage

		Probuphine (N=84)	SL BPN (N=89)	Total (N=173)
Number of Subjects who were dispensed supplemental SL BPN		15 (17.9%)	13 (14.6%)	28 (16.2%)
Total Number of Dispensing Episodes	1	5 (6.0%)	0 (0.0%)	5 (2.9%)
	2	2 (2.4%)	3 (3.4%)	5 (2.9%)
	3	0 (0.0%)	2 (2.2%)	2 (1.2%)
	4	1 (1.2%)	4 (4.5%)	5 (2.9%)
	5	2 (2.4%)	2 (2.2%)	4 (2.3%)
	6	3 (3.6%)	1 (1.1 %)	4 (1.2%)
	7	1 (1.2%)	1 (1.1%)	2 (1.2%)
	21	1 (1.2%)	0 (0.0%)	1 (0.6%)
Average Number of Tablets Dispensed and not Returned		42.9	24.9	34.5

SL BPN = sublingual buprenorphine

Figure 4 shows the distribution of the urine toxicology results over the entire study with the dispensing dates for the supplemental sublingual buprenorphine represented by blue circles. The blue lines indicate the estimated duration of supplemental buprenorphine use for each subject. The subjects above the black line either provided three positive urines or required supplemental buprenorphine and so could be non-responders. It is noted that a number of patients who required supplemental medication did not have positive test results. These patients could be considered adequately-treated if the dose of medication could be readily adjusted in response to patient need, which is possible with sublingual buprenorphine but not with Probuphine.

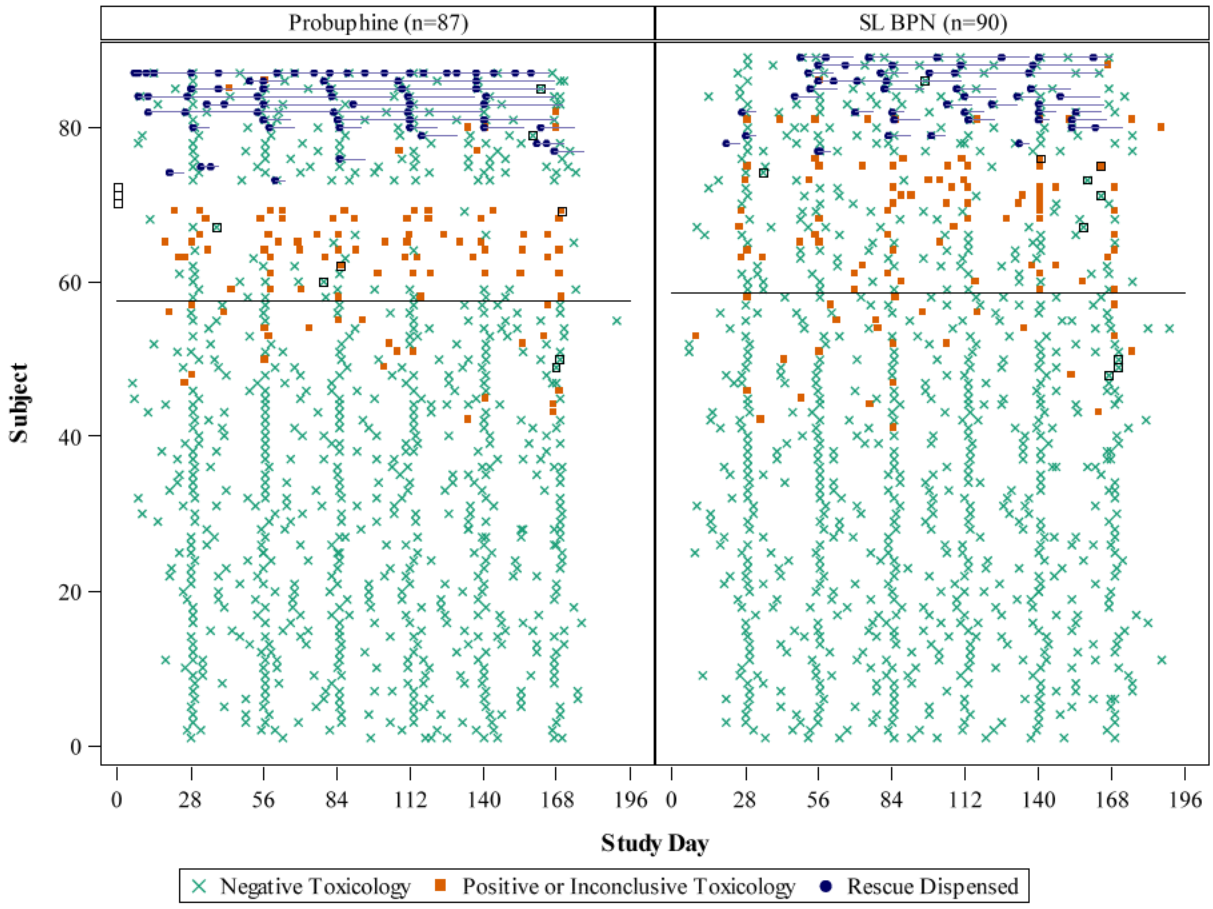
Figure 4: Urine Toxicology Results with Rescue Dispensing Dates with Missing Urine Tests



SL BPN = sublingual buprenorphine

Figure 5 shows the urine toxicology results with all visits where there were missing opioid panels indicated as positive.

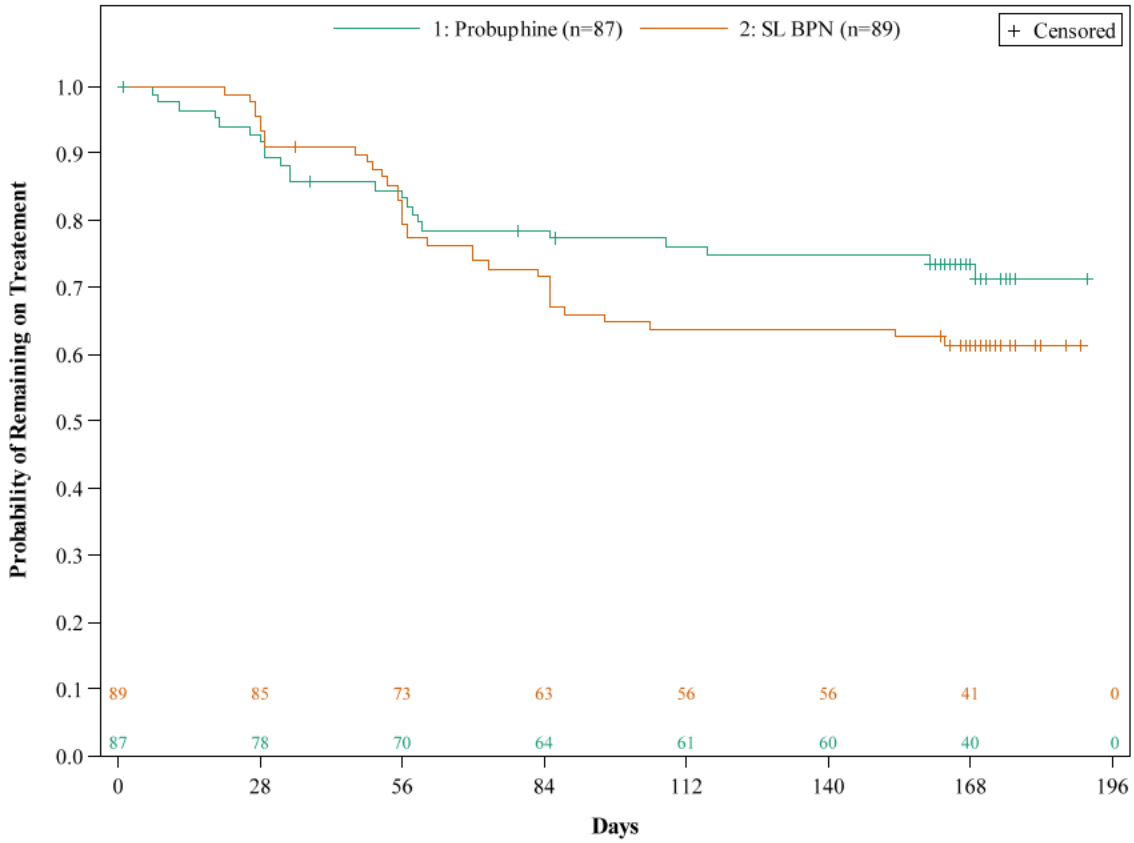
Figure 5: Results of Urine Tests with Missing Opioid Panels imputed as Positive



SL BPN = sublingual buprenorphine

Figure 6 shows the Kaplan-Meier analysis of the time to the first rescue, positive urine, or self-report of illicit opioid use. Note that only the urine toxicology tests reported as positive are included. The majority of first time rescue or illicit opioid use occurred by the end of the second month for the subjects in the Probuphine treatment group.

Figure 6: Kaplan-Meier Plot of Time to First Illicit Opioid Use or Supplemental Medication Dispensing



SL BPN = sublingual buprenorphine

Finally, to examine the impact of additional sublingual buprenorphine usage, those subjects who required additional sublingual buprenorphine were considered non-responders. Table 12 shows the percentage of subjects that were considered responders when missing urines samples imputed as positive and subjects that required supplemental sublingual buprenorphine were considered non-responders. An additional analysis was conducted were missing urines panels were also considered positive.

Table 12: Analysis of Supplemental Sublingual Buprenorphine Use

Category	Probuphine n (%)	SL BPN n (%)	Proportion Difference (95% CI) Probuphine – SL BPN	*P* Value (2-Sided)
Missing Urines imputed as Positive and Subjects with Sublingual Buprenorphine Use counted as Non-Responders				
N	87	89		
Responder	63 (72%)	65 (73%)	-0.006 (-0.138, 0.125)	0.93
Non-responder	24 (28%)	24 (27%)		
Missing Urines Panels imputed as Positive and Subjects with Sublingual Buprenorphine Use counted as Non-Responders				
N	87	89		
Responder	58 (67%)	59 (66%)	0.004 (-0.136, 0.143)	0.96
Non-responder	29 (33%)	30 (34%)		

SL BPN = sublingual buprenorphine

5.4.4 Success by Prior Dose

Probuphine delivers a fixed, non-adjustable dose of buprenorphine over a period of six months. Consequently, it is important to evaluate the difference in efficacy based on the prior dose that the subjects were receiving before the study as this may have an impact on the stability of the patient. Table 13 shows the breakdown of the subjects in the study by prior dose. The majority of the subjects in the study were receiving a dose of 8 mg per day prior to the study and approximately one-third of the subjects were receiving a sublingual film.

Table 13: Prior Dose by Study Treatment

Study Treatment	Prior Dose (mg)				
	2	4	6	8	Total
SL BPN	3 (3%)	15 (17%)	4 (4%)	67 (75%)	89
Probuphine	6 (7%)	12 (14%)	8 (9%)	61 (70%)	87

SL BPN = sublingual buprenorphine

Table 14 shows a summary of the success rates by prior dose. The number of subjects receiving less than 8 mg was relatively small and so these subjects were grouped together. The subjects receiving less than 8 mg were somewhat more likely to be classified as responders than subjects receiving 8 mg.

Table 14: Responder Rates by Prior Dose

Category	Probuphine n (%)	SL BPN n (%)
Subjects who Received 8 mg		
N	61	67
Responder	39 (64%)	42 (63%)
Non-responder	22 (36%)	25 (37%)
Subjects who Received less than 8 mg		
N	29	22
Responder	22 (76%)	17 (77%)
Non-responder	7 (24%)	5 (23%)

SL BPN = sublingual buprenorphine

5.4.5 Non-Inferiority Margin

The Applicant selected a non-inferiority margin of 20% for this study as it would preserve at least 70% of the expected effect of 75% for sublingual buprenorphine compared to placebo. (The expected effect is the difference between a predicted responder rate of 100% for sublingual buprenorphine and 25% for placebo.) There were two concerns noted with this expected effect. First, the observed responder rate for subjects receiving sublingual buprenorphine was far below the expected 100% and second, there is significant variability in the placebo response rate noted in the literature.

In Table 12 we present the results of several analyses that explore the effect of missing data and supplemental medication on the responder rates. The sublingual buprenorphine response rate for these analyses was 66% and 73% (two worst cases) and the lower bound of the 95% confidence interval for the difference between Probuphine and sublingual buprenorphine is approximately -0.14 (-14%) for both.

The estimated effect size of the sublingual buprenorphine, *C*, can be calculated by subtracting the hypothesized placebo response rate from the observed sublingual buprenorphine response rate. If we let *T* be the lower bound of the 95% confidence interval for the difference between Probuphine and sublingual buprenorphine, then we can compute a worst-case estimate of the proportion of the effect of sublingual buprenorphine that is preserved by Probuphine using the following formula:

$$Proportion\ Preserved = \frac{C+T}{C}$$

Table 15 shows the proportion of the estimated effect that is preserved for both these responder rates and for a range of placebo response rates. In all but one scenario, the proportion of the sublingual buprenorphine effect preserved is below 70%.

Table 15: Proportion of the Estimated Effect Size Preserved by Probuphine

Placebo Response Rate	SL BPN Response Rate			
	66%		73%	
	SL BPN Effect	% Preserved By Probuphine	SL BPN Effect	% Preserved By Probuphine
25%	41%	66%	48%	71%
30%	36%	61%	43%	67%
35%	31%	55%	38%	63%

SL BPN = sublingual buprenorphine

The Applicant chose to use 25% as their estimate of the placebo response rate for this study. This choice appears to have been based on the median expected responder rate found in the survey of addictions specialists reported in Table 6. However, the mean expected responder rate for this survey was 30%. There was also considerable variability among the relapse rates observed in the literature referenced by the Applicant. The applicability of these papers to the current study is however not clear due to the differences in the study population and study conduct. It's not clear what the best estimate of the placebo rate should be nor is it clear how much of the estimated effect should be preserved, i.e. what the correct NI margin should be.

5.5 Discussion

The results based on the protocol-specified analysis demonstrated that the proportion of responders among patients blindly switched to Probuphine was non-inferior to the proportion of responders who continued on sublingual buprenorphine. However, the responder rate depends on a number of assumptions about missing data and also assumes that use of supplemental buprenorphine is not an indicator of inadequacy of treatment. When analyzed under different assumptions, the response rates are lower than reported by the Applicant, and also differ from the expected response rate used to calculate the non-inferiority margin. Therefore, under some sets of assumptions, one might question whether enough of the effect size has been maintained to conclude efficacy of Probuphine. Moreover, because Probuphine ensures compliance, one would expect a clearer demonstration of superiority over sublingual buprenorphine than was demonstrated in this trial.

We will ask the committee to address whether the available efficacy data are sufficient to conclude that the drug is effective for the intended use. We will also ask whether the extent of efficacy demonstrated is sufficient to outweigh the risks, and whether further dose exploration should be required prior to approval.

6 Review of Safety

The safety evaluation of Probuphine centered on an assessment of the systemic effects of the active ingredient, buprenorphine, in the setting of continuous delivery of buprenorphine, provided with this formulation. The other principal focus was on the safety of Probuphine as it relates to the indwelling rods/implants and the insertion and removal procedures necessary for Probuphine administration.

The safety profile of the drug substance, buprenorphine, has been fairly well-characterized. Given that Probuphine provides lower levels of exposure to buprenorphine than the earlier transmucosal formulations, on which the safety profile is based, Probuphine could be expected to have a more favorable safety profile. Review of the Probuphine safety data did not identify major systemic safety concerns beyond those consistent with the established safety profile of buprenorphine. As such, a primary focus of this discussion will be on the formulation-specific safety findings and the procedural concerns unique to this novel buprenorphine delivery system (discussed in Section 6.1.5). Major safety results also will be summarized.

Probuphine Safety Database
The entire Probuphine safety database includes 647 unique subjects that participated in the PRO-814 trial, as well as the previously completed trials intended to support the original NDA submission. These included two Phase 3, six-month, placebo-controlled safety and efficacy trials (PRO-805 and PRO-806); two open-label extension studies (PRO-807 and PRO-811); a pharmacokinetic study (TTP- 400-02-01); and a comparative bioavailability study (PRO-810).

Safety Database – Original NDA Submission
A total of 450 opioid-dependent patients were enrolled in the phase 3 studies for the original NDA submission, of whom 222 received Probuphine implants and 109 received placebo implants; an additional 119 were treated with sublingual buprenorphine and received no implants. A subset of these patients continued into open-label extensions providing data for longer-term exposure. Including patients receiving Probuphine in safety studies after completing the placebo arm, 262 patients received Probuphine in these efficacy and safety studies.

Expanded Safety Database – NDA Resubmission
With the completion of the PRO-814 trial, an additional 176 patients were added to the safety database, of whom 87 received Probuphine and 89 received sublingual buprenorphine.

For the NDA Resubmission, the Applicant submitted an Integrated Summary of Safety (ISS) Addendum that pooled safety data from the PRO-814 trial with that of the other six-month controlled trials (PRO-806, PRO-806). Key differences between PRO-805 and

PRO-806, which were near identical in design, and the PRO-814 trial, are that PRO-805 and PRO-806 trials entailed the following trial design elements:

- enrolled new entrants to treatment who were required to reach a target dose of 12 to 16 mg/day of sublingual buprenorphine prior to insertion of implants
- had a placebo comparator arm and supplemental or rescue sublingual buprenorphine was permitted for these patients; and
- allowed for insertion of a fifth rod when certain protocol-specified criteria were met

PRO-806 also included an open-label (OL) sublingual buprenorphine arm. Subjects on sublingual buprenorphine in the PRO-806 OL arm were receiving 12–16 mg/day, approximately twice that of the patients in the PRO-814 arm who were receiving no more than 8 mg per day.

The resulting pooled safety database used for the ISS Addendum for this submission includes information for 626 subjects who had Probuphine or placebo implants or received sublingual buprenorphine in the pooled double-blind studies. The primary evaluation of safety is based on the pooled data from the three double-blind placebo- and/or active comparator-controlled trials. However, for deaths, nonfatal SAEs, withdrawals due to TEAEs, safety information from the OL studies (PRO-807 and PRO-811) as well as the clinical pharmacology studies (TTP-400-02-01 and PRO-810) is also summarized.

Note that, for the purposes of safety comparisons, patients identified as treated with "placebo," in most cases, received sublingual buprenorphine. Therefore, some common buprenorphine-associated adverse events may have occurred with similar frequency between the Probuphine arm and the placebo arm.

A summary of the 3 Phase 3 trials pooled for the safety database for the NDA Resubmission is provided in the following table.

Table 16: Phase 3 Trials Included in the Pooled Safety Database for the ISS Addendum

Study	Design	Treatment Groups		
		Probuphine	Placebo	SL BPN
Phase 3 Randomized, Controlled Studies				
PRO-805	Placebo-controlled, double-blind	108	55	--
PRO-806	Placebo-controlled, double-blind; open-label active-comparator control	114	54	119
PRO-814	Double-blind, double-dummy, active-controlled	87	--	89
	Treatment Arm Totals	309	109	208
	Combined Placebo/SL BPN Total		317	
	Phase 3 Safety Database Total		626	

SL BPN = sublingual buprenorphine

Again, as noted above, safety data from the PRO-814 trial were pooled only with the other two 24-week (six months) Phase 3 studies and were the focus of the integrated safety information for the NDA Resubmission. However, the complete safety database also includes safety data from two open-label trials and two clinical pharmacology trials.

Overall Exposure to Probuphine

The following provides a summary of overall cumulative exposure to Probuphine across all studies with Probuphine (including the open-label studies and the clinical pharmacology studies).

Table 17: Cumulative Exposure to Probuphine across All Probuphine Clinical Studies

	TTP-400-02-01	PRO-805/ PRO-807 Continuing Probuphine	PRO-807/ Placebo-> Probuphine	PRO-806/ PRO-811 Continuing Probuphine	PRO-811/ Placebo-> Probuphine	PRO-811/ SL BPN-> Probuphine	PRO-810	PRO-814/ Probuphine	Total
Any Exposure	12	108	12	114	8	20	9	87	370
≥1 month	12	102	12	112	8	18	8	84	356
≥2 months	12	99	12	109	7	17	0	83	339
≥3 months	12	96	11	108	7	16	0	81	331
≥4 months	12	87	11	103	7	14	0	81	315
≥5 months	12	81	10	99	7	14	0	81	304
≥6 months	0	60	3	78	2	7	0	1	151
≥7 months	0	47	0	54	0	0	0	0	101
≥8 months	0	42	0	53	0	0	0	0	95
≥9 months	0	41	0	52	0	0	0	0	93
≥10 months	0	41	0	50	0	0	0	0	91
≥11 months	0	39	0	46	0	0	0	0	85
≥12 months	0	39	0	46	0	0	0	0	85

Abbreviations: CRF = case report form; SL BPN = sublingual buprenorphine
Note: Study drug exposure is calculated as the date of implant removal minus the date on implant insertion plus one. If the removal date is missing, the study discontinuation date from the CRF is used. For PRO-805, PRO-806, PRO-807, and PRO-811, subjects who completed their respective study are assigned 6 months of exposure in that study. Subjects with more than 28 weeks of exposure in a given study are recorded as 28 weeks.
Note: A month is defined as 365.25/12 days.

Source: ISS Addendum.

Demographic and Baseline Characteristics in the Pooled Double-Blind Trials

Overall, the baseline characteristics of the populations were evenly distributed across the Probuphine arm and sublingual buprenorphine and placebo arms of the trials. Compared to patients in PRO-805 and PRO-806, which enrolled new entrants to treatment, subjects in PRO-814, who were considered "clinically stable," were more likely to report a prescription opioid reliever as their primary opioid of abuse (74.4%) as compared with 63% reporting heroin as the primary opioid of abuse in PRO-805 and PRO-806. Subjects in PRO-814 were a few years older on average, included fewer males, had higher BMIs on average, and were almost exclusively white.

Table 18: displays the demographic characteristics of the participants in the double-blind, controlled trials.

Table 18: Demographic and Baseline Characteristics, PRO-805, PRO-806, PRO-814

Demographic and Baseline Characteristics	Probuphine N=309 n (%)	Placebo/SL BPN N=317 n (%)	Total N=626 n (%)
Sex			
Male	196 (63)	195 (62)	391 (63)
Female	113 (37)	122 (39)	235 (38)
Race			
White	259 (84)	267 (84)	526 (84)
Black	31 (10)	31 (10)	62 (10)
Asian	1 (<1)	3 (1)	4 (<1)
American Indian or Alaskan Native	9 (3)	1 (<1)	10 (2)
Native Hawaiian or Other Pacific Islander	1 (<1)	--	1 (<1)
Other	8 (3)	15 (5)	23 (4)
Ethnicity			
Hispanic or Latino	39 (13)	43 (14)	82 (13)
Not Hispanic or Latino	270 (87)	274 (87)	544 (87)
Age (years)			
n	309	317	626
Mean (SD)	36.7 (11.06)	37.1 (11.04)	36.9 (11.04)
Median	35.0	35.0	35.0
Min, Max	19.0, 63.0	18.0, 64.0	18.0, 64.0
Age Groups			
18-35	156 (51)	166 (52)	322 (51)
36-65	153 (50)	151 (48)	304 (49)
Primary Opioid of Abuse			
Heroin	160 (52)	159 (50)	319 (51)
Prescription opioid pain medication	143 (46)	156 (49)	299 (48)
Other	5 (2)	2 (<1)	7 (1)
Opioid abuse treatment history			
Yes	230 (74)	226 (71)	456 (73)
No	79 (26)	88 (28)	167 (27)
Missing	--	3 (1)	3 (1)
BMI (kg/m^2)			
n	305	313	618
Mean (SD)	26.5 (5.99)	25.9 (5.58)	26.2 (5.79)
Median	25.5	24.9	25.2
Min, Max	13.6, 67.0	17.4, 54.7	13.6, 67.0
BMI group			
≤25 (kg/m^2)	141 (46)	162 (51)	303 (48)
>25 (kg/m^2)	168 (54)	155 (49)	323 (52)

SL BPN = sublingual buprenorphine

6.1 Major Safety Results

6.1.1 Deaths

There were no deaths in Probuphine-treated patients. One death occurred in the sublingual buprenorphine arm in Study PRO-806, attributed to heroin overdose.

6.1.2 Serious Adverse Events

Serious adverse events (SAEs) were reported in 10 (3%) of the patients randomized to Probuphine in the controlled trials, 7 (6%) of the patients randomized to placebo, and 9 (4%) of the patients randomized to sublingual buprenorphine. Additionally, 3 SAEs were reported in patients continuing on Probuphine in the open-label extensions and in one patient who completed placebo treatment in the controlled studies and was started on Probuphine in the open-label extension. Several of the events were of an infectious nature, including abscesses potentially related to intravenous drug use. Depression and suicidal ideation were also reported. The pattern of SAEs did not identify novel systemic findings inconsistent with the known safety profile of buprenorphine. One SAE related to the implant site was reported in a patient who received a placebo implant. However, because the risks of insertion are likely to be related to the procedure, and not to the drug, this event is of concern even in a placebo-treated patient. Table 19 briefly lists the types of events reported.

Table 19: Serious Adverse Events, Probuphine Clinical Studies

Pooled DB Studies – PRO805,PRO-806, & PRO-814		
Probuphine (3%)	**Placebo (6%)**	**SL BPN (4%)**
1. Hypotension (sepsis, BP meds) (806) 2. COPD exacerbation & Pulmonary Embolism (805) 3. Pneumococcal pneumonia (806) 4. Pneumonia (806) 5. Umbilical hernia, obstructive (806) 6. Tooth abscess (806) 7. Second degree burns (805) 8. Breast cancer (806) 9. Convulsion (814) 10. Bipolar I disorder (814)	1. Respiratory Failure (heroin withdrawal; agitation → intubation) (805) 2. Suicidal Ideation((805) 3. Tylenol overdose (per pt, 50 500-mg pills) (806) 4. Pneumonia; Explant Site Cellulitis (805) 5. Limb abscess (not implant site) (806) 6. Gastroenteritis (806) 7. Relapse (805)	1. Major depression (806) 2. Pyrexia (infxn) (806) 3. Angina pectoris (806) 4. Pulmonary embolism (806) 5. Rib fracture (806) 6. Spontaneous abortion (806) 7. Biliary colic (814) 8. Chronic cholecystitis (814) 9. Bronchitis (814)
Pooled OL Extension Studies – PRO-807 & PRO-811		
Probuphine → Probuphine	**Placebo → Probuphine**	
1. CAD, worsening 2. Cellulitis antecubital fossa (methamphetamine IV) 3. Suicidal ideation	1. Pneumonia (pneumonia in primary study also)	
Clinical Pharmacology Study PRO-810		
1. pancreatic cyst x 2, nausea		

SL BPN = sublingual buprenorphine
PR0-805 treatment arms included Probuphine and placebo.
PRO-806 treatment arms included Probuphine, placebo, and an OL SL buprenorphine (BPN).
PRO-814 treatment arms included Probuphine and SL buprenorphine.
Supplemental or rescue buprenorphine was permitted for all the treatment arms.

6.1.3 Adverse Events Leading to Discontinuation

Adverse events leading to discontinuation were uncommon in both active- and placebo-treated patients. Notably, the most common type of event leading to discontinuation, involved problems at the implant site. However, all of these occurred in Study PRO-805 and its extension, Study PRO-807. The procedures used for insertion/removal in those studies differed from those used in PRO-806, PRO-811, and PRO-814, which were conducted after the applicator underwent refinement (a bevel-tipped applicator replaced the blunt-tipped applicator used in PRO-805 and PRO-807), and the procedures for insertion and removal were also improved. Table 20 illustrates events leading to discontinuation, noting the study number in which the event was reported.

Table 20: Adverse Events Leading to Patient Discontinuation

Pooled Double-Blind Studies – PRO-805,PRO-806, & PRO-814		
Probuphine (3%)	**Placebo (2%)**	**SL BPN (4%)**
1. Implant Site Pain/Infection (805) 2. Implant Site Pain/Infection (805) 3. Implant Site Pain (805) 4. Hepatic Enzyme increases (805) 5. LFT abnormal (806) 6. Breast Cancer (806) 7. Muscle Spasms [↑ CPK] (814)	1. Tylenol Overdose (806) 2. Hepatitis C (806)	1. ALT/AST increased (806) 2. Weight decreased (806) 3. Neck pain (806) 4. Sinus tachycardia (806) 5. Drug dependence (806)
Pooled OL Extension Studies – PRO-807 & PRO-811		
1. Implant Site Hemorrhage, Infection, Edema & Erythema (807) – Probuphine 2. Implant Site Infection (807) – Probuphine 3. ALT increased (811) Probuphine → SL BPN induction		
Clinical Pharmacology Study PRO-810		
1. Adverse dropout in BA Study (810) – pancreatic cyst		

SL BPN = sublingual buprenorphine
PR0-805 treatment arms included Probuphine and placebo.
PRO-806 treatment arms included Probuphine, placebo, and an OL SL buprenorphine (BPN).
PRO-814 treatment arms included Probuphine and SL buprenorphine.
Supplemental or rescue buprenorphine was permitted for all the treatment arms.

6.1.4 Common Adverse Events:

The adverse event profile of sublingual buprenorphine has been previously characterized in the safety database for buprenorphine sublingual tablets and buprenorphine/naloxone sublingual tablets. The adverse event tables from the approved labeling are shown in Appendix D.

The table below illustrates the common adverse events in the pooled double-blind studies in the Probuphine development program, excluding those events related to the implant site or insertion and removal procedures. (These events are discussed separately in Section 6.1.5.1.)

Table 21: Common Non-Implant Site Treatment-Emergent Adverse Events (≥ 2% in the Probuphine group or Placebo/ Sublingual Buprenorphine group) in the Pooled Double-Blind Studies, PRO-805, PRO-806 and PRO-814

MedDRA Preferred Term	Probuphine N=309 n (%)	Placebo/SL BPN N=317 n (%)	Total N=626 n (%)
Any Non-Implant Site TEAE	200 (64.7)	205 (64.7)	405 (64.7)
Headache	39 (12.6)	32 (10.1)	71 (11.3)
Insomnia	26 (8.4)	36 (11.4)	62 (9.9)
Nasopharyngitis	27 (8.7)	22 (6.9)	49 (7.8)
Upper respiratory tract infection	25 (8.1)	23 (7.3)	48 (7.7)
Nausea	20 (6.5)	15 (4.7)	35 (5.6)
Anxiety	15 (4.9)	18 (5.7)	33 (5.3)
Back pain	18 (5.8)	15 (4.7)	33 (5.3)
Depression	20 (6.5)	10 (3.2)	30 (4.8)
Constipation	20 (6.5)	9 (2.8)	29 (4.6)
Vomiting	17 (5.5)	11 (3.5)	28 (4.5)
Toothache	14 (4.5)	10 (3.2)	24 (3.8)
Oropharyngeal pain	14 (4.5)	10 (3.2)	24 (3.8)
Diarrhoea	10 (3.2)	13 (4.1)	23 (3.7)
Pain	12 (3.9)	9 (2.8)	21 (3.4)
Dizziness	11 (3.6)	7 (2.2)	18 (2.9)
Abdominal pain upper	10 (3.2)	7 (2.2)	17 (2.7)
Gastroenteritis viral	7 (2.3)	9 (2.8)	16 (2.6)
Influenza	7 (2.3)	9 (2.8)	16 (2.6)
Bronchitis	6 (1.9)	9 (2.8)	15 (2.4)
Abscess limb	4 (1.3)	10 (3.2)	14 (2.2)
Tooth abscess	6 (1.9)	8 (2.5)	14 (2.2)
Contusion	6 (1.9)	8 (2.5)	14 (2.2)
Cough	10 (3.2)	4 (1.3)	14 (2.2)
Hyperhidrosis	8 (2.6)	6 (1.9)	14 (2.2)
Urinary tract infection	7 (2.3)	6 (1.9)	13 (2.1)
Fatigue	9 (2.9)	4 (1.3)	13 (2.1)
Arthralgia	6 (1.9)	7 (2.2)	13 (2.1)

Abbreviations: SL BPN = sublingual buprenorphine; TEAE = treatment-emergent adverse event

The common adverse events observed with Probuphine, buprenorphine in an implantable delivery system, and with sublingual buprenorphine are reported with similar frequencies. A few adverse events were reported somewhat more frequently in the Probuphine arm, including depression.

6.1.5 AEs of Special Interest

As described earlier, the systemic safety of buprenorphine has been well-characterized. The Probuphine safety database did not reveal new systemic safety concerns overall for buprenorphine in this novel delivery system. As a result, the safety evaluation focused on matters pertaining specifically to the implantable delivery system, including safety issues associated with the indwelling rods/implants and the surgical procedures for insertion and removal. Adverse events of special interest for this NDA Resubmission thus include implant site reactions and insertion and/or removal complications. Findings related to hepatic effects and QT prolongation will also be discussed in brief. Review of the findings in the Probuphine safety database pertaining to these latter safety concerns has determined that the understanding of the risk-benefit profile of buprenorphine in transmucosal forms has not been altered.

The following three adverse events of special interest will be discussed in this section, with the greater part of the discussion focused on the first item.

1. Implant site reactions and complications of insertion or removal
2. Hepatic effects
3. QT prolongation

6.1.5.1 Implant site reactions and complications of insertion or removal

As mentioned previously in this briefing document, Probuphine requires a minor surgical procedure for insertion of the rods or implants, and another procedure for removal of the rods six months later. Over the course of the Probuphine development program, the Applicant made modifications to the equipment, to the design of the applicator, and to the general surgical techniques related to the insertion and removal of the implants, most notably, adopting the "U-technique" for removal. The training procedures also underwent modification. These modifications were implemented in an effort to improve safety outcomes related to the indwelling rods/implants and procedures for insertion and removal, as described in documentation provided by the Applicant.

The Applicant implemented the Probuphine Clinical Training and Certification program and the applicator was modified to (1) ensure that clinicians who performed the implant insertion and removal procedures met competency standards and (2) improve the overall safety of the implant and removal processes. The original applicator in the PRO-805 and PRO-807 studies was blunt-tipped and was associated with more tissue adhesions to the implants, resulting in more implant fractures. The Applicant explains that the modified sharp applicator mitigated tissue damage and allowed closer placement of implants, thus facilitating easier removals.

The Applicant also performed Human Factors testing and validation of the proposed training program. The Applicant conducted a human factors validation of the Probuphine training program and associated instructional materials, with the primary objective of the effort being to validate a single Probuphine training program's effectiveness in preparing intended users to perform the Probuphine insertion and removal procedures.

Obstetrician/Gynecology medical experts in the Division of Bone, Reproductive, and Urologic Products (DBRUP) were asked to review the procedural-related safety findings and to provide a clinical perspective on the Human Factors Study and findings, drawing from their knowledge of and experience with implantable contraceptives.

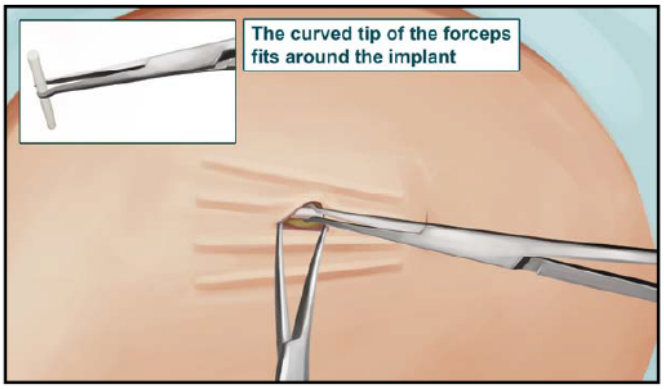
6.1.5.1.1 Proceduralist Perspective on Procedural-Related Safety and Human Factors Study for Probuphine Insertion and Removal Procedures

Issues related to the procedures for insertion and removal were reviewed by consultants in the Division of Bone, Reproductive, and Urologic Products (DBRUP). In reviewing the safety experience from the clinical trials, DBRUP noted that the extent of procedural exposure is small relative to the pre-approval exposure for the contraceptive implants. For example, prior to approval, the Norplant clinical program included 849 removal procedures. The Jadelle (2 levonorgestrel implants) program had > 1100 removal procedures, whereas the Implanon (etonogestrel implants) and Nexplanon (radiopaque version of Implanon) programs had 942 and 296 removal procedures, respectively.[22] DBRUP also noted that generalizability of adverse event profiles of contraceptive implants to Probuphine implants may be inferred if Probuphine providers have reasonably similar surgical expertise as providers of contraceptive implants.

The text below is largely excerpted from DBRUP's consult.

> Among the concerns relating to the insertion/removal procedures identified in the first review cycle was the "U-technique" used for removal of the buprenorphine rods (see [Figure 7 below]). This technique is not commonly practiced in the US, and its adoption was a subject of questions raised by gynecology experts at the 2013 Advisory Committee meeting.

Figure 7: Incision for removal of Probuphine implants



Source: Figure 16, Probuphine Instruction for Insertion and Removal, Attachment B to the Response to FDA Information Request, dated November 10, 2015

[22] DBRUP slide presentation to the 2013 Probuphine Advisory Committee.

[DBRUP determined that the U-technique] was originally described in 1993 by Dr. Untung Praptohardjo for Norplant removal, and subsequently modified by Reynolds in 1995.[23,24] In this technique, a 4 mm incision was made longitudinally between capsules 3 and 4, starting approximately 0.5 cm proximal to the distal ends of the capsules, rather than transversely at the base of the capsules. Forceps were inserted through the incision to grasp the Norplant capsule at right angles to its long axis and within 5 mm of the distal tip. The capsule was pulled to the incision, while the handle of the forceps was rotated toward the subject's shoulder, bringing the tip of the capsule into view in the incision. The fibrous capsule was cleaned off and the capsule was removed. This technique was shown to shorten removal times and was associated with less damage to the implants. Reynolds made minor modifications to the U-technique so the implants could be grasped anywhere along the shaft.

DBRUP noted the following with respect to the U-technique. Compared to the U-technique, the Applicant's training material describes a larger incision (7-10 mm) to be made, necessitating suturing for wound closure. The Applicant has noted that the Probuphine implants appear to be "less forgiving" than Norplant implants; the larger incision proposed likely allows greater potential for the ease of dissection and better access to the implant in the event of implant breakage.

It should be noted that, if the patient plans to continue with another treatment cycle with Probuphine, a separate incision would be needed for insertion at this visit. This is in contrast with contraceptive implants, where new implants, if requested by patients, are usually inserted through the same incision made for removal in the opposite direction to the implants previously placed. In DBRUP's assessment, the U-technique likely provides greater visualization of and access to the implants to facilitate removal and poses little additional risk.

6.1.5.1.1.1 Review of Procedure-Related Safety -- Clinical Trials in NDA 204442

DBRUP reviewed procedure-related safety issues based on information provided in the Integrated Summary of Safety relating to the following clinical studies:

- 2 double-blind, placebo-controlled trials (Study PRO-805 and Study PRO-806) and 1 double-blind, active-controlled trial (PRO-814) where subjects receiving the active control product (sublingual buprenorphine) also received placebo implants
- 2 open-label, safety extension studies (PRO-807 and PRO-811)
- 1 open-label, comparative bioavailability study (PRO-810)
- 1 dose-finding pharmacokinetic study (TTP-100-02-01)

[23] Praptohardjo U, Wibowo S. The "U" technique: a new method for Norplant@ implants removal. Contraception 1993;48:526-536.
[24] Reynolds RD. The "modified U" technique: a refined method of Norplant removal. J Fam Pract. 1995;40(2):173-80.

Of note, several modifications were made during development to the applicator used for the insertion of Probuphine implants, the number of implants inserted, and the insertion/removal procedures. DBRUP's safety review has taken these changes into account. With respect to the applicator, its original blunt-tip, which was associated with more encapsulation and implant fractures, was changed to a sharp one to reduce tissue damage, allow for closer placement of implants and easier removals. Additionally, for Studies PRO-806 (henceforth referred to as "806" for brevity), 811, and 814, the Probuphine Clinical Training and Certification program was implemented. Clinicians watched an Implant Insertion/Removal Training Video and were given written instructions for the proper, aseptic subdermal insertion and removal of Probuphine and placebo implants. Finally, subjects in the first two efficacy trials (Studies 805 and 806) initially received 4 implants but were allowed to receive a 1-implant dose increase (arriving at 5 implants total) if protocol dose increase criteria were met. All subjects in the third trial (Study 814) received 4 implants (either Probuphine or placebo); a fifth implant was not permitted.

To assess procedure-related safety, DBRUP pooled procedures performed across five trials – three efficacy trials (805, 806, 814) and two extension trials (807 and 811) in which subjects received a second treatment cycle. Cumulative exposure to the insertion/removal procedures among subjects who participated in these five trials is shown in **Table** 22] below:

Table 22. Pooled Extent of Exposure to Procedures

Number of subjects	Probuphine implants	Placebo implants	Total
Study 805	108	55	163
Study 806	114	54	168
Study 814	87	89	176
Study 807	62	N/A	62
Study 811	85	N/A	85
	456	198	654

Source: Extracted from Table 5, disposition of Subjects by Study, pages 31-32 of 153, ISS Addendum, Module 5.3.5.3; NDA 204442/0000: Table 10-1Disposition of Subjects (safety population) Clinical Study Report, page 65, Study Report Body PRO-807, Module 5.3.5.2; 204442/0000: Table 10-1Disposition of Subjects (safety population) Clinical Study Report, page 65, Study Report Body, Module 5.3.5.2204442/0000: Table 10-1Disposition of Subjects (safety population) Clinical Study Report, page 65, Study Report Body PRO-811, Module 5.3.5.2, page 65.

As expected, commonly reported procedure-related adverse events (AEs) were mild and self-limiting, such as pain, pruritis, erythema at the incision/implant site. Procedure-related AEs of special interest are summarized in [Table 23] below. Compared to contraceptive implants, higher incidences of bleeding (10.9%), complicated removals (3.2%), and implant site infection (4.0%) were noted in the Probuphine trials.

Of note, DBRUP disagreed with the Applicant's categorization of AEs associated with "complication of device removal." In the Applicant's individual study reports and the integrated safety summary, subjects who required a second

attempt to remove all implants were not deemed to have "complicated removal." DBRUP considered a failure to remove all implants during the first attempt – thus necessitating imaging studies to locate all implants and a second removal attempt – to be a complication of the initial implant removal attempt.

Table 23: Key Procedure-Related Adverse Events by Trial

	Efficacy Studies			Extension Studies			
	Study 805 (N = 163)	Study 806 (N = 168)	Study 814 (N = 176)	Study 807 (N = 62)	Study 811 (N = 85)	Total # Events of Special Interest	AE incidence (% of Total # Procedures Performed, 654)
Implant expulsion[ɤ]	4 (2.5%)	1 (0.6%)	1 (0.6%)	2 (4.8%)	0	8	1.2%
Implant site infection*	9 (5.5%)	3 (1.8%)	6 (3.4%)	4 (6.4%)	4 (4.7%)	26	4.0%
Wound complications[∞]	4 (2.5%)	2 (1.2%)	2 (1.1%)	1 (1.6%)	1 (1.1%)	10	1.5%
Complicated removal or requiring 2nd attempt	15 (9.2%)	0	7 (4%)	3 (4.8%)	2 (2.3%)	27	4.1%
Bleeding**	30 (18.4%)	19 (11.3%)	1 (0.6%)	16 (25.8%)	5 (5.9%)	71	10.9%

Source:

- For Study 805: extracted from Table 15/page 78 of Study Report, Table 2 and written response to Information Request dated 2/28/13
- For Study 806: extracted from Table 14.3.1.2 of Study Report, response to Information Request dated 2/28/13
- For Study 814: extracted from Table 30 of Study Report
- For Study 807: extracted from Table 14.3.1.2.1 of Study Report, response to Information Request dated 2/28/13
- For Study 811: extracted from Table 14.3.1.2 of Study Report, response to Information Request dated 2/28/13
- For Study 814: extracted from Table 30 of Study Report

ɤ including implant expulsion and implant protrusion
*including AE terms of cellulitis, purulent discharge, implant site pruritus, incision site infection, and wound infection, implant site abscess, and subcutaneous abscess
∞ including AE terms of incision site necrosis, wound dehiscence, incision site complication, postoperative wound complication, suture-related complication, wound complication, impaired healing
**including AE terms of implant site bleeding/hematoma/hemorrhage, and incision site hemorrhage

DBRUP explained that Key implant site AEs in the database fall into the following broad categories:

- *Pain*
- *Hemorrhage/hematoma*
- *Infection (includes general term infection, cellulitis, wound infection)*
- *Device expulsion*
- *Complicated removal*
- *Neuropathy (paresthesias, peripheral sensory neuropathy)*

[DBRUP's] review did not identify any long-term complications such as permanent disability due to nerve damage; it would be unlikely for such events to be observed in a clinical program of this size.

Two events types of special interest emerged from [DBRUP'] review of procedure-related safety. First, pooled incidences of bleeding in the Probuphine program, including implant site hemorrhage/hematoma and incision site bleeding (10.9%) is much higher than that (of hematoma) observed in the Implanon clinical program (0.1%).[25] *Second, implant site infections were seen at a relatively high rate for a simple procedure in the setting of subdermal implant insertion (4.0% overall).*

Two explanations may be plausible for such observations. First, the general health status of patients with addiction likely differs from that of young, generally healthy women who seek long-term contraceptive implants. Thus, greater AE incidences would be expected (and likely unavoidable) in the Probuphine program. Two, not all providers who performed the procedures in the Probuphine program were equally familiar with surgical care (both intra-operative/technical care and postoperative care); it is conceivable that providers who were less procedurally-oriented may have had worse surgical outcomes. For example, rates of hematoma and hemorrhage were higher when the procedures were performed by psychiatrists and family medicine practitioners than for surgical specialists in the Studies 805 and 806.[26] *It is possible that competency in pre-operative procedures, insertion and removal of such implants is expected to improve over time given sufficient surgical volume and continuing education/training. However, if Probuphine is approved, these safety findings suggest that provider qualification and training should be better defined. Furthermore, continued provider training and enhanced pharmacovigilance for procedure-related AEs should be considered.*

6.1.5.1.1.2 Clinical Perspective--Human Factors Study

In [a] 2013 consult review, DBRUP summarized the profile of more serious safety concerns associated with contraceptive implants, including:

- Complicated removal due to deep placement or broken implants
- Migration of existing implants including to other sites in the arm or chest
- Nerve damage (from either deep placement or complicated removal), potentially resulting in permanent disability
- Partially removed implants, possibly due to encapsulation from fibrous tissue

[25] Implant label, section 6.1. http://www.accessdata fda.gov/drugsatfda_docs/label/2015/021529s011lbl.pdf
[26] NDA 204442 Integrated Summary of Safety Attachment A Tables 1-4, pages 1-8/8.

- Inability to locate implants for removal, necessitating additional invasive surgery
- Infection
- Bleeding
- Spontaneous expulsion

In response to the 2013 Complete Response letter, the Applicant developed training materials and Instruction for Use to mitigate these potential risks. The Applicant conducted a series of human factors reviews and formative studies (collectively submitted as the "Human Factor study"), seeking to validate the effectiveness of a single training program in preparing healthcare providers to perform Probuphine insertion and removal procedures.

Components of the Human Factor Study include:

1. Formative study: Classroom instruction on implant procedures: slide presentation on the anatomy of the brachium, the insertion procedure, implant localization, removal procedure, wound care and voiding complications. The moderator instructed participants to review the Instruction for use (IFU) and view videos of both insertion and removal procedures. The participants then performed the procedures without assistance from trainers.
2. Effectiveness training: Live practicum of procedures using a simulated human arm (i.e. pork tenderloin), focus on proper techniques to avoid complications
    a. To simulate the removal procedure, each piece of pork tenderloin had 4 placebo implants placed 1-4 prior to the practicum. One implant was intentionally fractured (into two pieces of equal size). Another implant had adhesive injected around it to simulate adherent/fibrotic tissue that would require dissection
3. Certification exam: Each participant was evaluated on implant insertion and removal procedure performance, and on responses to a series of knowledge-based questions on both insertion and removal procedures.
    a. Metrics used to evaluate performance include:
        i. Insertion Procedure
            1. Maintaining a sterile field
            2. Proper incision performance
            3. Proper Probuphine applicator usage
            4. Implant depth
            5. Implant distribution
        ii. Removal Procedure
            1. Identification of all four implants--with or without imaging assistance (Ultrasound or MRI)
            2. Maintaining a sterile field
            3. Proper incision performance
            4. Proper dissection technique (if necessary)

b. Critical tasks and subtasks which may mitigate potential risks that were evaluated pertaining to patient screening, insertion, removal, and patient discharge.

The Applicant recruited both "proceduralists" and "non-proceduralists" to participate, as intended providers of Probuphine did not appear to be limited to only providers with surgical expertise. The human factor study qualified physicians and mid-level providers as "proceduralists" if they meet one of these two criteria:

- They had completed a medical residency or fellowship in a "procedural specialty" AND they currently practiced in that specialty. (A "procedural specialty" was defined as one in which practitioners perform invasive procedures involving injection of local anesthetic and use of sterile technique, including but are not limited to: anesthesia, surgery, obstetrics and gynecology, dermatology, emergency medicine, critical care, etc.)
- They had performed a sterile procedure in the last 3 months, defined as injecting local anesthetic AND using sterile technique to place sutures, insert a catheter, or make a skin incision. If a midlevel provider, nurse practitioners and physician assistants only.

Both proceduralists and non-proceduralists participated in the classroom instruction and formative user testing (using pork tenderloins) to assess the number of successful implant completion and implant depth. However, the Applicant subsequently allowed only proceduralists to participate in the live practicum/certification portion of the human factor study. The live practicum portion enrolled 15 proceduralists with diverse backgrounds – physicians from multiple specialties (anesthesia, surgery, obstetrics and gynecology, dermatology, emergency medicine, critical care, etc.) as well as midlevel providers (nurse practitioners and physician assistants.

Reviewer comment:
The metrics, critical tasks and subtasks are adequate to capture deficiencies in preventing the AEs of concern. However, DBRUP has the following concerns with the overall design, and in turn the utility, of human factor study:

- *The pork tenderloin may be suitable as a model for demonstrating technical proficiency for the insertion procedure. However, it is not suitable for predicting whether certain procedure-related AEs - such as infection and bleeding – can be mitigated by training. As a consequence, the only pertinent task that can be assessed was "depth of implant placement," which on its own has limited clinical relevance.*
- *The scenarios designed to mimic complicated removals (from either breakage or densely adhesed implants) appeared reasonable.*

However, the pork tenderloin is not adequate as a substitute for the removal procedure. Neither the pork tenderloin nor an artificial arm can provide an adequate representation of scarring after 6 months of foreign bodies in the arm. In addition, neither substitute would allow for real-world scenarios in which patients may move, experience pain requiring more anesthesia, or have bleeding.

- *The Applicant has not clearly articulated who the intended real-world "proceduralists" would be, but participants in the simulation/validation component of human factors study were all from specialties which involve doing procedures or surgery. Consequently, results of this human factor study are not generalizable to providers of other non-surgical specialties. If approved, DBRUP recommends that labeling and risk mitigation and evaluation strategies (REMS) specify the qualification of the providers who will be performing the insertion/removal procedures.*
- *The Applicant should require mid-level providers to also be licensed and provide experience of "procedural specialty" as in many states mid-level providers work independently from physicians.*

Results of human factor study

Results of training with IFU/video viewing showed that physicians (both proceduralists and non-proceduralists) performed slightly better than mid-level practitioners (both non-proceduralists and non-proceduralists), as shown in [Table 24 and Table 25] below.

Table 24: Implant Depth and Distribution Correctness by Subgroup

User Subgroups	Implant 1 Depth Correctness	Implant 2 Depth Correctness	Implant 3 Depth Correctness	Implant 4 Depth Correctness	Distribution Correctness
Proceduralist - Physicians	7 of 8	7 of 8	6 of 8	5 of 8	6 of 8
Proceduralist - Midlevels	5 of 7	5 of 7	5 of 7	5 of 7	4 of 7
Nonproceduralist - Physicians	6 of 7	6 of 7	6 of 7	5 of 7	5 of 7
Nonproceduralist - Midlevels	5 of 8	5 of 8	5 of 8	4 of 8	3 of 8

Table 25: Implant Removal Performance

User Group	Incision Length Correctness	Incision Depth Correctness	Implants Removed and/or Appropriate Imaging Referral Initiated
Proceduralist - Physicians	5 of 8	7 of 8	4 of 8
Proceduralist - Midlevels	7 of 7	7 of 7	3 of 7
Nonproceduralist - Physicians	5 of 7	7 of 7	5 of 7
Nonproceduralist - Midlevels	4 of 8	5 of 8	2 of 8

The Applicant acknowledged that "when users were provided with only the IFU and video materials to…prepare themselves for performing the insertion and removal of Probuphine, there was sub-optimal performance." However, it is unclear why they proceeded to the live practicum/validation portion of the study only with proceduralists (both physicians and mid-level practitioners) as non-proceduralist physicians appeared to have performed better than proceduralists-midlevel practitioners.

Results of the live practicum/validation study are shown in [Table 26] below. For the purposes of this review, this medical officer grouped salient subtasks according to the AE to be mitigated.

Table 26: Risks and Subtasks to Mitigate These Risks

Risk	Subtask	Insertion	Removal	Correctly Performed
Infection	Using aseptic technique, place applicator and four implants on the sterile field	Subtask #11		13/15 (corrected* to 15/15)
	Clean incision site area with chloraprep triple swab for up to 30 seconds	Subtask #13	Subtask #56	14/15 (corrected* 15/15)
	Unwrap surgical tray and place equipment in the sterile field		Subtask #54	14/15 (corrected* 15/15)
Deep placement(which could result in migration/ spontaneous expulsion/nerve damage/hemorrhage or hematoma)	Check applicator function by removing the obturator from the cannula and re-locking it	Subtask #12		13/15 (corrected* 15/15)
	While tenting, gently advance applicator			12/15 x 4 =48; 6 too shallow (<5mm); 6 too deep (5-7 mm) Corrected* 15/15 (60/60)
Bleeding	Make a 2.5-3mm length shallow incision at the marked insertion site	Subtask #17		15/15
Lost migrated implants Difficult/Incomplete Removal/Broken implant	Locate non-palpable implants with ultrasound or MRI		Subtask #48	14/15 (corrected* 15/15)
	Make a 7-10mm incision w/ the scalpel, parallel to the access of the arm, between the 2nd and third implants		Subtask #60	14/15 (midlevel providers made incisions 20-22 mm long)
	Lift skin edge with Adson toothed forceps		Subtask #61	14/15 (used mosquito clamp), corrected* 15/15
Fibrous encapsulation	Dissect away any tissue adhering to the implant w/ scissors or mosquito forceps		Subtask #62	14/15 (one closed incision without removing difficult implant; would send for imaging first)
	If implant is encapsulated, use scalpel to shave tissue sheath and carefully dissect the implant		Subtask #64	

After the live practicum, participants were asked follow-up questions to assess their performance and knowledge. The Applicant concluded that the participants performed well on assigned tasks. However, a closer reading of the narratives yielded the following gaps in participants' responses, which DBRUP considers notable for having potential clinical ramifications:

On mitigating infection risks:

- 14/15 participants succeeded in inserting all 4 implants. One participant reported getting "flustered" and placed the 4th implant outside the sterile zone.[27]
- 14/15 participants cleaned incision site with antiseptic prior to insertion. One participant omitted this step despite acknowledging this instruction in training.[28]
- 14/15 participants properly placed sterile equipment on the sterile field. One participant broke sterile field while wearing a nonsterile glove despite knowing the importance of properly maintained sterile field.[29]

On mitigating the risks resulting from complicated removal:

- 7/15 participants succeeded in removing all 4 implants. The other 8 "followed proper safety protocol."[30]
- 1 mid-level practitioner was unable to remove all 4 implants but proceeded to close the incision. This participant indicated that it "would be prudent to close the incision, bandage up, and send the patient for imaging to return 2-3 weeks later).[31]
- 14/15 participants correctly requested imaging studies (ultrasound or MRI) to locate non-palpable implants prior to making an incision for removal. A mid-level practitioner failed to request imaging prior to making an incision despite indicating that imaging would have been warranted. She indicated that "she would have followed that guideline in a real patient situation." The Applicant interpreted her response as "correct" due to "study artifact."[32]

On proper use of instruments and surgical technique:

- 13/15 successfully confirmed that the applicator was functioning properly before initiating the procedure. Because all participants indicated that functionality of the applicator should be checked and the "low likelihood of applicator malfunction," the Applicant stated that knowledge "was transferred adequately" and that the two violations of this task "were not due to the training program deficiency."[33]
- 12/15 participants correctly tented up the skin while advancing the applicator in the 60 implant attempts (4 per participants). As a result, some implants were placed either too shallow (6 of 12 attempts) or too deep (6 of 12 attempts) relative to the pre-specified and recommended subdermal

[27] Page 55 of 193, human factor study report.
[28] Page 57-58 of 193, human factor study repot.
[29] Page 57 of 193, human factor study report.
[30] Page 61 of 193, human factor study report.
[31] Page 59 of 193, human factor study report.
[32] Page 57 of 193, human factor study report.
[33] Page 50-51 of 193, human factor study report.

level. The applicant attributed these as "slips" (presumably, from tenting).[34]

Reviewer comment:
Given the design of this human study, the subtasks and critical tasks identified appear appropriate. The study showed that most participants could adequately perform the tasks required to mitigate the risks of infection, bleeding and fibrous scar formation around implants. Nevertheless, the narratives of task failures captured above raise a number of issues:

- *The Applicant appears to equate "receipt of knowledge" with the ability to adequately perform a surgical procedure. It is unclear how "transfer of knowledge" can mitigate the procedure-related safety concerns that were identified in the clinical trials. The Applicant appears to assume, that once a provider recognizes their task failure, they would be able to perform this task correctly in subsequent procedures. However, by design, the human factor study provides no data to support such an assumption.*
- *There were three task failures relating to mitigating infection risks in this human factor study. Notably, the overall incidence of infection-related AEs (4.0%, of all procedures performed) in the clinical trials were already high for an outpatient procedure, aseptic technique and maintaining sterile field should be further addressed in the training program if Probuphine is approved.*
- *Not all participants were able to remove all implants in this practice session. The Applicant has not adequately articulated how complicated removals —which will include non-localized, deep or broken implants— will be addressed in the real world setting. Based on postmarketing data on contraceptive implants, implants have been known to migrate great distances from the site of insertion. The Applicant should have a plan for localizing Probuphine implants that are not found with ultrasound or MRI of the upper arm. Further, postmarketing data indicate some contraceptive implants are never localized or removed. The Applicant should address follow-up if implants are never localized or removed.*
- W*ith regard to deep insertion, 6 of 60 (10%) of implants inserted were beyond the desired depth (5-7 mm); some implants were appropriately positioned and some too deep in any given insertion of 4. All of the deep placements were by midlevel providers. None reached or exceeded the depth of 10 mm which the Applicant associates with a risk of acute or chronic injury to a patient. While DBRUP concurs that insertion depth less than 10 mm is unlikely to result in injury, the finding suggests that the steps in the training program related to insertion depth should be reinforced.*

[34] Page 53-54 of 193, human factor study report.

6.1.5.1.2 Expulsions and Extrusions

During the Probuphine clinical trials, the implant site was to be evaluated at each clinic visit. The implant site was to be visually inspected for evidence of erythema, edema, itching, pain, infection, bleeding, abnormal healing, and any other abnormalities. The implant site was also to be examined for evidence of removal or attempted removal of the implants.

Several patients experienced complications such as complete expulsion and protrusion of extrusion of implants, in addition to the complications described earlier in this discussion of procedural-related safety. These cases are described below.

Implant Expulsions

All expulsions of the implants occurred in the Probuphine arms of the trials.

- 002-019 – 20M presented to clinic 1 week post the Week4 visit, pulled out 3 Probuphine implants at visit, and 4th was removed by a study physician. The patient admitted to attempting to remove the protruding rods at home, prior to presenting to clinic, and was discontinued from study. (PRO-805)

- 608-025 – 36M reported that one implant had "popped through" while the patient showered approximately one month after insertion. The patient brought in the implant. The patient had an infection at the implant site, was to have a replacement implant, but infection continued. The patient was arrested for a probation violation, and an additional implant came out at that time that he threw away. Infection resolved, remaining 2 implants removed, U/S performed to confirm. (PRO-806)

- 021-001 – 51F had a protruding implant in PRO-805 about one month after insertion and implant replacement, had 2 broken implants about two months after insertion replaced with 2 new implants. All 5 implants were removed without incident at the end of study (there was an option for placement of a fifth implant in this trial). In PRO-807, Probuphine implants replaced on 3 separate occasions due to various implant site TEAEs.

Implant Extrusions (Protrusion)

Implant extrusions and protrusions primarily occurred in the Probuphine arms of the trials. A single case of an implant protrusion occurred in PRO-814.

- 006-003 – 40M had "implant fragment surfacing" approximately 7 months after insertion. Pt recovered, no action taken regarding implant. (PRO-805)

- 021-001 – 51F, as above for (PRO-805). Patient experienced both expulsions and protrusions.

- 027-013 – 27M had 2 implants "extruding from incision site" approximately 3 weeks after insertion, and 2 new implants were inserted. Multiple implant site reactions during the timeframe. (PRO-805).

- 004-001 – 27F, "implant site extrusion of 2 implants," approximately 2.5 months after insertion. No action was taken, and subject recovered (PRO-807).

- 010-003 – 57 M in PRO-814 developed cellulitis at the implant site on Day 5, treated with oral antibiotics, and cellulitis reported as resolved on Day 18. On Day 30, protrusion of 1 rod without complete expulsion was observed and was reported as implant site TEAE of expulsion of implant. On Day 31, the subject had his 4 implants removed from the left arm and 4 new placebo implants were inserted in his right arm.

Although Probuphine provides some safeguard against abuse and misuse because it is an implantable formulation, the rods/implants have the potential to protrude or to be completely expelled from the skin either spontaneously or intentionally. As such, the risk of risk of abuse, misuse, and accidental exposure is not eliminated with Probuphine, and in the context of the expulsions and protrusions, there is a potential for these public health risks to be realized.

6.1.5.2 Hepatic Effects

Buprenorphine has been associated with hepatitis and other hepatic events. The *Warnings and Precautions* section of current labeling for sublingual buprenorphine (as Suboxone) includes safety labeling regarding hepatitis and hepatic events as follows:

5.6 Hepatitis, Hepatic Events

Cases of cytolytic hepatitis and hepatitis with jaundice have been observed in individuals receiving buprenorphine in clinical trials and through post-marketing adverse event reports. The spectrum of abnormalities ranges from transient asymptomatic elevations in hepatic transaminases to case reports of death, hepatic failure, hepatic necrosis, hepatorenal syndrome, and hepatic encephalopathy. In many cases, the presence of pre-existing liver enzyme abnormalities, infection with hepatitis B or hepatitis C virus, concomitant usage of other potentially hepatotoxic drugs, and ongoing injecting drug use may have played a causative or contributory role. In other cases, insufficient data were available to determine the etiology of the abnormality. Withdrawal of buprenorphine has resulted in amelioration of acute hepatitis in some cases; however, in other cases no dose reduction was necessary. The possibility exists that buprenorphine had a causative or contributory role in the development of the hepatic abnormality in some cases. Liver function tests, prior to initiation of treatment is recommended to establish a baseline. Periodic monitoring of liver function during treatment is also recommended. A biological and etiological evaluation is recommended when a hepatic event is suspected. Depending on the case, SUBOXONE sublingual tablet may need to be carefully discontinued to prevent withdrawal signs and symptoms and a return by the patient to illicit drug use, and strict monitoring of the patient should be initiated.

The Applicant reported that "Hy's Law cases"[35] cases, which are considered indicative of potential drug-induced liver injury, were identified in the PRO-805 and PRO-806 controlled trials. The Applicant previously identified cases in the open-label extensions also. However, the Applicant applied an additional total bilirubin (TBL) criteria of >50% elevated over baseline, beyond the customary TBL >2xULN criterion for bilirubin, and identified Hy's Law cases based on this broader definition, rather than applying only the customary TBL >2xULN criterion for bilirubin. No cases of simultaneous AST or ALT elevations above 3xULN and total bilirubin elevations above 2xULN were identified in the Probuphine safety database.

The Probuphine safety database reveals no new hepatic safety concerns beyond those previously identified in the clinical trial and postmarketing setting for the marketed sublingual buprenorphine products.

6.1.5.3 QT prolongation

A signal for QT prolongation has been identified in a study of transdermal buprenorphine used for analgesia. The extent of prolongation noted was considered to meet the

[35] Hy's Law cases have the following three components:

1. The drug causes hepatocellular injury, generally shown by a higher incidence of 3-fold or greater elevations above the ULN of ALT or AST than the (nonhepatotoxic) control drug or placebo
2. Among trial subjects showing such AT elevations, often with ATs much greater than 3xULN, one or more also show elevation of serum TBL to >2xULN, without initial findings of cholestasis (elevated serum ALP)
3. No other reason can be found to explain the combination of increased AT and TBL, such as viral hepatitis A, B, or C; preexisting or acute liver disease; or another drug capable of causing the observed injury

Excerpt from Guidance for Industry Drug-Induced Liver Injury: Premarketing Clinical Evaluation available at: http://www.fda.gov/downloads/Drugs/.../Guidances/UCM174090.pdf

threshold for regulatory concern, a value which is used to determine whether or not the effect of a drug on the QT/QTc interval in target patient populations should be studied intensively during later stages of drug development. The potential for doses of buprenorphine used for the treatment of opioid dependence to prolong the QT interval has not yet been evaluated in formal QT studies.

Electrocardiogram (ECG) data were evaluated for the pooled double-blind and open-label studies, and elevations above baseline have been noted. The data confirm that QT prolongation may be seen in patients treated with buprenorphine.

The Applicant has been notified that a post-marketing requirement would be imposed to conduct a trial to assess the risk of QT prolongation with subdermal buprenorphine.

6.2 Safety Summary

In general, the common adverse events associated with Probuphine treatment were similar to those seen with transmucosal forms of buprenorphine treatment. The hepatic effects and effects on cardiac conduction were also consistent with buprenorphine's expected effects. The most notable adverse events for Probuphine were related to the rods/implants, indwelling foreign bodies, and to the surgical procedures related to insertion and removal.

In a safety database comprising 626 subjects who participated in the Phase 3 program, 7 (1%) experienced expulsions or extrusions of implants. Five patients discontinued treatment due to implant-site adverse events. All but one of the patients who experienced expulsions, extrusions, or AEs leading to discontinuation was on Probuphine, suggesting that the irritancy of buprenorphine could play a role in these implant site adverse events. More minor implant-site AEs were reported in a significant number of study participants, even after implementation of a modified insertion device and training procedure. The incidence of complicated removals and other adverse events including hematoma and infection are noteworthy, and are important for an understanding of the safety of Probuphine and in defining risk mitigation strategies.

7 Discussion and Points for Consideration

The original studies raised concern about the adequacy of the plasma level of buprenorphine delivered by Probuphine for the treatment of a broad population of individuals with opioid dependence. In the newly-conducted study, the Applicant attempted to identify a patient population for whom the dose provided by Probuphine would be adequate. This population was defined, for the purposes of study entry criteria, by both clinical stability and pre-study sublingual buprenorphine dose.

We will ask the Committee to discuss whether the Applicant has succeeded in identifying a population for whom Probuphine is effective. This will involve discussing two major issues:

1. Does the submitted study provide evidence of efficacy for treatment with Probuphine in the studied population? Or is further dose exploration for this population needed?

2. If there is evidence to support efficacy of Probuphine in a specific population, what factors define a patient who would be a candidate for this treatment?

First, we noted that this is a novel study design, in a population not usually enrolled in addiction treatment trials, using infrequent visits and urine toxicology testing, and defining a responder based on a combination of self-report and urine toxicology findings. It was also designed as a non-inferiority trial, which rests on a number of assumptions.

The Applicant reports very high responder rates in both treatment arms. As described in the background document, the agreed-upon protocol-specified definition of responder did not include use of supplemental sublingual buprenorphine because it was not anticipated that stable patients who had been on a particular dose for a substantial period of time without drug use would require dose adjustments. However, the results showed use of rescue in 16% of patients (15% of patients on sublingual buprenorphine, 18% of patients on Probuphine). Some patients were successfully managed with dose adjustments, as evidenced by a lack of self-reported or urine toxicology-detected drug use, but in clinical practice, Probuphine-treated patients would not necessarily be seen for regular visits with supplemental buprenorphine dose adjustments. We will ask the Committee to discuss under what circumstances should patients requiring rescue be adjudicated, for the purposes of analysis of the efficacy of Probuphine, as non-responders, given the non-titratable nature of the product. Would this be based on dose required, number of occasions, time until rescue, or other factors?

The protocol did not pre-specify how urine samples that were collected, but not properly analyzed, would be handled in the efficacy analysis. The sampling schedule was infrequent compared to customary efficacy studies in this indication (monthly vs. as much as thrice-weekly), and only 10 samples per patient were to be collected. Roughly 12% in each arm have samples entirely missing due to missed visits, but in addition, because of sample handling issues, 22% (25% in the Probuphine arm and 18% in the sublingual buprenorphine arm) are missing data from one or more of the samples. We will ask the Committee to discuss what assumptions should be made to handle missing data from urine toxicology samples, and whether the significance of missing samples would be different under different circumstances. For example, is there greater concern over samples that were not collected because the patient did not appear for a visit, compared to samples that were collected as scheduled, but not properly analyzed? For samples that were not collected, is there greater significance to a patient failing to appear to submit a random sample vs. failing to attend a scheduled visit vs. a sample that was refused, forgotten, or overlooked?

These issues are important because they will have bearing on how the results of the study, including the expected rate of response to Probuphine treatment, are communicated in

labeling and in promotion. Discuss which factors (use of rescue, missing random samples) you would include in identifying successful patients. Given the findings concerning rescue use and the nature of missing information, what is the best way to express the results of the study in terms of response rates?

The data on the extent of use of supplemental buprenorphine raise a question of whether Probuphine actually provides the purported advantage over sublingual buprenorphine with respect to diversion and accidental pediatric exposure. We will ask the Committee to discuss how the observed frequency of supplemental sublingual buprenorphine rescue could translate to clinical practice. If a certain amount of supplemental rescue buprenorphine is to be expected at some phase of treatment, should clinicians be advised to routinely prescribe a supply of "just-in-case" sublingual buprenorphine to patients receiving Probuphine? If so, discuss how this will impact the product's ability to mitigate misuse, abuse, and accidental pediatric exposure.

If, on the other hand, use of rescue should be interpreted as an indicator that the patient is not well-managed with Probuphine, should labeling advise providers not to continue Probuphine if rescue is routinely required? We will ask the Committee to comment on whether rescue at different times in treatment (shortly after insertion vs. the end of the dosing period vs. throughout the 6 months) are of similar significance, and how to communicate the approach to handling each of these scenarios.

We will also ask the Committee to discuss what patient selection criteria should be used to identify patients who are candidates for Probuphine treatment. The labeling proposed by the Applicant suggests that baseline sublingual buprenorphine dose is the only characteristic required.

The Applicant has provided information on a training and certification program to ensure that practitioners can safely insert Probuphine. However, because fibrotic tissue develops around the implants, the procedure of removing Probuphine after six months of insertion is not readily modeled for the purposes of training. Based on experience involving contraceptive implants, it is known that complicated removals may require imaging equipment and surgical exploration. We will ask the Committee to discuss concerns about the adequacy of the proposed program to ensure Probuphine will be inserted and removed safely.

The Applicant has proposed a Risk Evaluation and Mitigation Strategy (REMS) consisting of a training/certification program for healthcare professionals who will prescribe Probuphine and for healthcare providers who will insert or remove Probuphine. Additionally, the REMS will restrict distribution to REMS-certified prescribers. We will ask the Committee to discuss whether the proposed REMS is adequate to address the risks of potential complications associated with improper insertion and removal, as well as, abuse, misuse, and accidental overdose if an implant protrudes or completely comes out of the skin.

8 Appendices

Appendix A Drug Addiction Treatment Act of 2000

The Narcotic Addict Treatment Act of 1974 limits methadone maintenance treatment to the context of the Opioid Treatment Program (OTP) (i.e., methadone clinic) setting. Methadone treatment of opioid addiction is delivered in a closed distribution system that originally required special licensing by both Federal and State authorities. The current regulatory system is accreditation-based, but OTPs must still comply with specific regulations that pertain to the way clinics are run, the credentials of staff, and the delivery of care. To receive methadone maintenance, patients are required to attend an OTP, usually on a daily basis, with the possibility of earning the privilege of taking home doses as their treatment stability increases.

Because this is the setting where addiction treatment was delivered for decades, most U.S. physicians have little experience and expertise in the treatment of opioid addiction.

The Title XXXV of the Children's Health Act of 2000 (P.L. 106-310) provides a "Waiver Authority for Physicians Who Dispense or Prescribe Certain Narcotic Drugs for Maintenance Treatment or Detoxification Treatment of Opioid-Dependent Patients." This part of the law is known as the Drug Addiction Treatment Act of 2000 (DATA 2000). Under the provisions of DATA 2000, qualifying physicians may obtain a waiver from the special registration requirements in the Narcotic Addict Treatment Act of 1974, and its enabling regulations, to treat opioid addiction with Schedule III, IV, and V opioid medications that have been specifically approved by FDA for that indication, and to prescribe and/or dispense these medications in treatment settings other than licensed OTPs, including in office-based settings. At present, the only products covered by DATA 2000 (i.e., Schedule III-IV, approved for the indication) are buprenorphine sublingual tablets and buprenorphine/naloxone sublingual tablets and films.

To qualify for a DATA 2000 waiver, physicians must have completed at least 8 hours of approved training in the treatment of opioid addiction or have certain other qualifications defined in the legislation (e.g., clinical research experience with the treatment medication, certification in addiction medicine) and must attest that they can provide or refer patients to necessary, concurrent psychosocial services. The 8 hour training courses are provided by various physician organizations (e.g. APA) and delivered in-person, in web-based formats, or through other mechanisms. Physicians who obtain DATA 2000 waivers may treat opioid addiction with products covered by the law in any appropriate clinical settings in which they are credentialed to practice medicine.

Appendix B Efficacy Results from Original NDA Submission

Background Related to Efficacy Endpoints and Study Design

A key issue in this application is the matter of "clinical significance" of the efficacy results. Addiction is a chronic, relapsing disorder in which patients self-administer drugs despite harmful consequences. There has been considerable debate about the proper endpoints to measure in clinical trials for addiction treatments, most of which have focused on attempts to quantify the use of the patients "drug-of-choice" and the effects that medications have on modifying that use.

However, ultimately, the goal of treatment is to produce a clinical benefit—a health benefit, or a benefit in terms of psychosocial or occupational functioning, or a mortality benefit—through suppression (or elimination) of drug use. Drug-taking behavior itself, observed during the brief window of a clinical trial, is a surrogate endpoint. Trials intended to show effects on physical or psychosocial consequences of drug use would need to be very long and very large, and may be impractical. However, when drug-taking behavior is used as a surrogate endpoint, there should be a demonstration of change in behavior that can be reasonably predictive of improvement, such as avoidance of drug-related health and social consequences. Trials demonstrating that patients attain and sustain abstinence from drug use have always been considered to provide compelling evidence of efficacy, without requiring direct measure of clinical benefit (e.g., without validation of abstinence as a surrogate for clinical benefit). Validation of other patterns of behavior as surrogates for clinical benefit can be accomplished by examination of data on long-term functioning of treated individuals comparing use patterns with outcomes—this has been accomplished to validate an endpoint short of abstinence for alcoholism treatment, for example. However, no such validation of other patterns of behavior as predictors of clinical benefit has been undertaken for opiate addiction.

Previous trials for medications to treat opiate addiction have used a variety of measures, including group mean proportion of opioid-negative urine samples, retention in treatment, longest period of abstinence, or other measures. There has not been a consensus on how to approach this problem.

In the development program for Probuphine, the Applicant was advised that analyses focused on group means (such as mean percent of weeks abstinent) were difficult to interpret, because they do not reflect the experience of individual patients, who might range from complete responders to non-responders. In light of this ambiguity, the appropriate endpoints and analytic approach were the subject of considerable debate over the course of the development program.

The emphasis was on trying to define a successful patient in such a way that patients who were clearly clinically successful would not be misclassified as unsuccessful due to a too-stringent definition, such as complete abstinence and attendance at all visits. It is

understood that some patients might be fully successful and yet miss some treatment visits, or might achieve full abstinence, but not by the end of a protocol-specified month or two of grace, or even that a fully-successful patient might "slip up" on occasion. The discussions did not contemplate that patients who achieved only minimal reductions in their drug use could or should be classified as successful.
The original NDA included two placebo-controlled clinical trials, PRO-805 and PRO-806. Both were randomized, double-blind, parallel-group, multi-center studies involving efficacy ascertainment over 24 weeks after insertion of Probuphine or placebo implants. The study designs were essentially identical, except that PRO-806 also included a treatment group in which patients were treated with open-label sublingual buprenorphine.

Eligible participants included patients 18 to 65 years of age who met DSM-IV criteria for current opioid dependence and had not received treatment in the past 90 days. Patients were to undergo initiation of buprenorphine treatment (induction) using sublingual tablets. In order to be randomized to treatment with Probuphine or placebo implant, patients had to meet the following criteria[36] after the induction phase:

- Completed induction with sublingual buprenorphine to a dose of 12–16 mg/day as clinically appropriate within 10 days. Patients requiring <12 mg/day or >16 mg/day were ineligible.
- No significant withdrawal symptoms (defined as a score ≤ 12 on the Clinical Opiate Withdrawal Scale [COWS])
- No significant cravings for opioids (defined as a score ≤ 20-mm on the 100-mm Opioid Craving Visual Analog Scale [VAS])

Insertion of Probuphine occurred within 12 to 24 hours after the last dose of sublingual buprenorphine. Patients were treated for 24 weeks on study. Following the randomization visit, there were approximately 88 scheduled visits: 16 study visits and 72 urine collection visits.

The protocols allowed for administration of supplemental sublingual buprenorphine during the study for symptoms of withdrawal or "craving" or on request at the discretion of the investigator[37]. Investigators were blind to the urine toxicology results; therefore, *supplemental buprenorphine was not provided on the basis of ongoing illicit drug use.*

[36] A substantial number of patients screened for inclusion failed to meet these criteria. Waivers were granted for some patients who needed additional time to stabilize or when implantation could not be scheduled in the designated window for logistical reasons. However, for a significant number of screen failures, the reason cited was that the patient was not able to be stabilized on a dose of 12-16 mg over three consecutive days within the specified window.

[37] Criteria for supplemental sublingual buprenorphine were:

- Withdrawal symptoms scoring >12 on COWS
- Request for dose increase by subject that was considered appropriate by investigator
- Cravings >20 mm on the Opioid Craving VAS

In Study 805, patients needed to meet only one criterion to receive rescue medication; in Study 806, patients needed to meet *all three* criteria.

Each dose of supplemental sublingual buprenorphine could only be obtained by patients at their clinic or pharmacy. Take-home sublingual buprenorphine was allowed for weekends, holidays, or other circumstances at the discretion of the investigator. Subjects in the open-label sublingual buprenorphine arm in Study PRO-806 could be provided up to seven days' supply of sublingual buprenorphine at a time.

Treatment failure was defined as

- Requiring supplemental sublingual buprenorphine exceeding the following limits, after having received the optional 5th implant[38]:
  - ≥3 days per week for 2 consecutive weeks
  - ≥8 days over 4 consecutive weeks at any time after the implant dose increase
- Requiring >1 additional day per week of counseling for 4 consecutive weeks (i.e., >3 sessions per week during Weeks 1 through 12 and >2 sessions per week during Weeks 13 through 24)

(Note: results of urine testing for opioid use were not included in criteria for treatment failure or in the criteria for rescue use. Therefore, patients could engage in ongoing illicit drug use without being adjudicated as treatment failures if they did not manifest signs of withdrawal, report craving, or request rescue.)

Any subject who requested, or who met one or more of the following criteria was withdrawn from the study:

- Subject non-compliance, defined as refusal or inability to adhere to the study protocol
  - missing 9 consecutive urine collections after the baseline visit
  - missing 6 consecutive counseling sessions after the baseline visit
  - refusal or inability to adhere to the study protocol, as determined by the principal investigator
- Evidence of implant removal or attempted implant removal
- Unacceptable or intolerable treatment-related AE
- Pregnancy
- Use of other treatments for opioid dependence
- Use of any investigational treatment
- Intercurrent illness or circumstances (e.g., incarceration ≥7 days) that, in the judgment of the investigator, affected assessments of clinical status to a significant extent
- Requirement for continual use of opioid analgesics >7 days or general anesthesia for surgery
- Lost to follow-up
- Treatment failure, as defined above

[38] After the first two weeks, if a subject met criteria for supplemental sublingual buprenorphine dosing on 3 or more days per week for 2 consecutive weeks or on 8 or more days total over 4 consecutive weeks, the subject received an implant dose increase.

Any subject who met the above criteria was seen for an end of treatment visit (unless lost to follow-up), during which implants were removed and clinical evaluations performed.

The insertion procedure was performed by a health care provider who had received training from the Applicant on the technique. For Study PRO-805, the training consisted of a DVD and self-teaching materials. New training procedures and a new insertion device was developed after completion of Study 805, and for Study PRO-806, in-person training using an improved device was instituted. Additionally, a somewhat novel approach to removing the implants was employed, using an incision that ran parallel to the implants rather than a perpendicular incision near the insertion incision. New pieces of equipment were provided to facilitate removal via this alternate method. Insertion and removal procedures were typically provided by a specific "implanting physician" at each site. At some sites, the general management of the patient's addiction problem was handled by one individual (e.g., in the Department of Psychiatry) and arrangements were made for a physician with surgical experience (e.g., in the Department of Gynecology) to perform the insertion and removal procedures.

The primary efficacy outcome for both studies was the cumulative distribution function (CDF) of the percent of urine samples negative for opioids.[39] The endpoint of interest for both studies was the CDF of the percentage of negative urines for Weeks 1 – 24 with self-report imputation. This endpoint was based on urine toxicology findings. Urine samples were taken three times per week during the studies, and tested for opioids with the exception of buprenorphine, as well as other illicit drugs.
A total of 331 patients were randomized to treatment with Probuphine (n = 222) or placebo (n = 109) in Studies PRO-805 and PRO-806.

- In Study PRO-805, 348 patients were screened and 163 were randomized in a 2:1 ratio to either Probuphine or placebo. This study was conducted at 23 sites in the United States. The first patient was enrolled on April 2, 2007, and the study was completed on June 19, 2008.

- Study PRO-806, 480 patients were screened and 287 were randomized in a 2:2:1 ratio to either Probuphine, open-label sublingual buprenorphine 12-16 mg per day, or placebo. The study was conducted at 20 sites in the United States. The

[39] Study PRO-805 was the first Phase 3 trial in the clinical development program, and the CDFs were based on negative urine samples during Weeks 1 through 16. When the Applicant entered Phase 3 of the development program, the Applicant still had some uncertainty about the full duration of therapy with the implant. While the Applicant was operating under the theory that the implant provided buprenorphine for a total of six months, they acknowledged that it was conceivable that it only delivered active drug for four months. For statistical reasons, the four-month window was designated the primary analysis and the six-month window, secondary. Since they judged that the implant lasts for six months, it renders the fourth month evaluations irrelevant, notwithstanding its identification as "primary" in the protocol.

first patient was enrolled on April 22, 2010, and the study was completed on May 12, 2011.

In general, the subject population across the two trials primarily consisted of White, non-Hispanic males in their mid-thirties who used heroin as their primary opioid of abuse and had received treatment for opioid abuse in the past. Most patients had been diagnosed within the five years preceding entry into the study. In PRO-806, a slightly higher proportion of females were enrolled, and the percentage of subjects with previous treatment history was smaller.

Patient disposition is illustrated below. Overall, 35% of the Probuphine-treated patients and 72% of the placebo-treated patients in the controlled trials did not complete the full 24 weeks of treatment. In the placebo arms, the most common reason for premature discontinuation was "treatment failure," which, again, was defined as requiring more than the protocol-specified limit of supplemental sublingual buprenorphine. Continued use of illicit substances was not considered in the definition of treatment failure, nor was continued use of illicit substances a criterion for receiving rescue medication. Based on the criteria for rescue medication, "treatment failure" refers specifically to inadequacy of treatment of patient-reported symptoms of withdrawal and "craving." The differences in the protocols with respect to providing rescue (one criteria needed to be met in PRO-805 while all three criteria for rescue had to be met to receive rescue in PRO-806) are reflected in the different rates of "treatment failure" in the placebo group between the two studies. Higher rates of "subject non-compliance" may have reflected dissatisfaction with placebo treatment with strict rescue criteria. High rates of loss to follow-up in the open-label sublingual buprenorphine arm may have reflected the fact that patients could access buprenorphine treatment with a less burdensome visit schedule outside of the study.

The table below includes both PRO-805 and PRO-806, and their respective open-label extensions, PRO-807 and PRO-811.

Patient Disposition Phase 3 Efficacy Studies and Safety Extensions

	Double-Blind Studies					Open-Label Studies	
	Study PRO-805		Study PRO-806			Study PRO-807	Study PRO-811
Disposition	Probuphine N=108 n (%)	Placebo N=55 n (%)	Probuphine N=114 n (%)	Placebo N=54 n (%)	SL BPN N=119 n (%)	Probuphine N=62 n (%)	Probuphine N=85 n (%)
Subject Completed Study	71 (65.7)	17 (30.9)	73 (64.0)	14 (25.9)	76 (63.9)	46 (74.2)	67 (78.8)
Subject Withdrew Early	37 (34.3)	38 (69.1)	41 (36.0)	40 (74.1)	43 (36.1)	16 (25.8)	18 (21.2)
Most Common Reasons for Early Withdrawal							
Subject Request	8 (7.4)	9 (16.4)	5 (4.4)	9 (16.7)	4 (3.4)	5 (8.1)	7 (8.2)
Subject Non-Compliance	12 (11.1)	7 (12.7)	10 (8.8)	9 (16.7)	8 (6.7)	5 (8.1)	0 (0.0)
Treatment Failure	0 (0.0)	17 (30.9)	6 (5.3)	9 (16.7)	0 (0.0)	0 (0.0)	1 (1.2)
Unacceptable or intolerable treatment-related adverse event	4 (3.7)	0 (0.0)	0 (0.0)	0 (0.0)	1 (0.8)	2 (3.2)	0 (0.0)
Intercurrent illness or circumstance that affected assessments of clinical status or required discontinuation of drug or both	1 (0.9)	0 (0.0)	8 (7.0)	4 (7.4)	8 (6.7)	0 (0.0)	3 (3.5)
Lost to Follow-Up	10 (9.3)	4 (7.3)	9 (7.9)	3 (5.6)	17 (14.3)	4 (6.5)	6 (7.1)

0

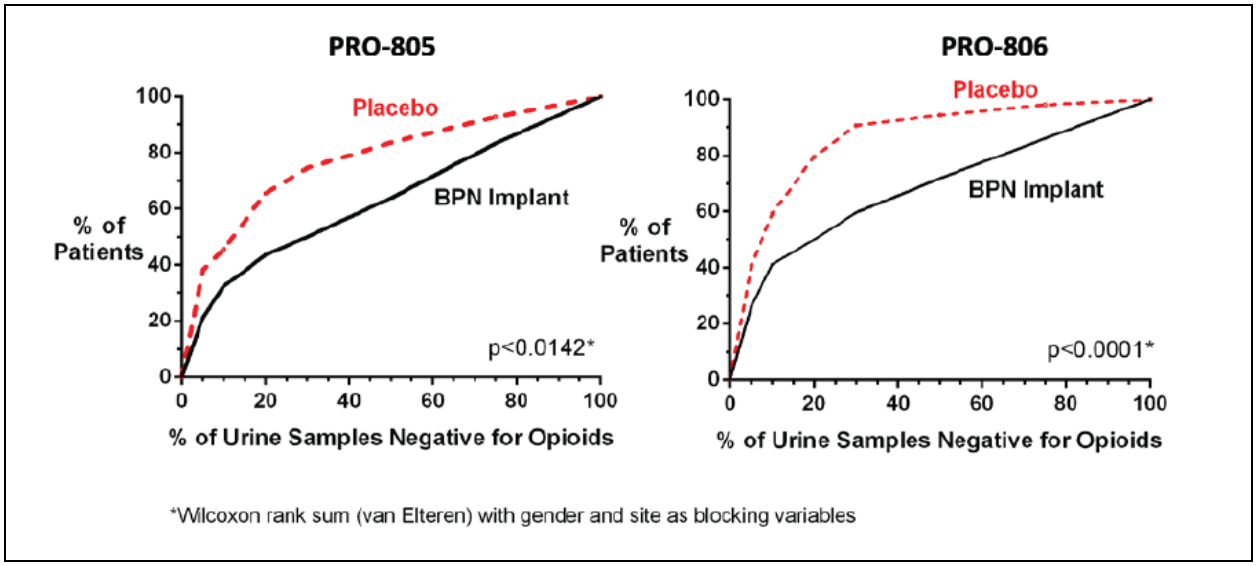
Abbreviations BPN = buprenorphine; SL = sublingual
Note: Percent for each reason for early withdrawal is based on the total number of subjects in the population.
Source: Summary of Clinical Safety, Table 7, page 48.

The primary efficacy analysis compared the cumulative distribution function (CDF) of the percentage of urine samples negative for opioids in the two treatment groups using a stratified Wilcoxon rank sum test with pooled site and gender as stratification variables.

The primary analysis for both studies was conducted by Biostatistics Reviewer, David Petullo, M.S., on the intent-to-treat population, defined as all randomized patients who received an implant. The percentage of negative urines was derived for each patient by summing the total number of negative urine samples and dividing by all possible samples. For weeks 1-24, the denominator was 72. For some patients, the denominator was greater as they had unscheduled urine test results. Missing samples were considered positive. If a patient reported illicit use of opioids during a specific week, urine samples collected during that timeframe were considered positive even if a urine sample tested negative. All results presented below were obtained by incorporating self-reported use.

The Applicant's graphic representations of the study results are presented in the following two figures.

Cumulative Distribution Function of the Percentage of Urine Samples Negative for Opioids in Weeks 1–24, with Imputation for Patient Illicit Opioid Self-Report: Studies PRO-805 and PRO-806



Source: Figure 15, Applicant's Advisory Committee Backgrounder

In the Applicant's presentations, the data are shown in graphs that illustrate the proportion of patients who submitted a particular percentage of opioid-negative tests *or fewer.* Although there is nothing technically or statistically wrong with these presentations, they are difficult to interpret intuitively. They can be compared to a survival curve that graphs how many patients died on a particular day *or sooner.* Like the

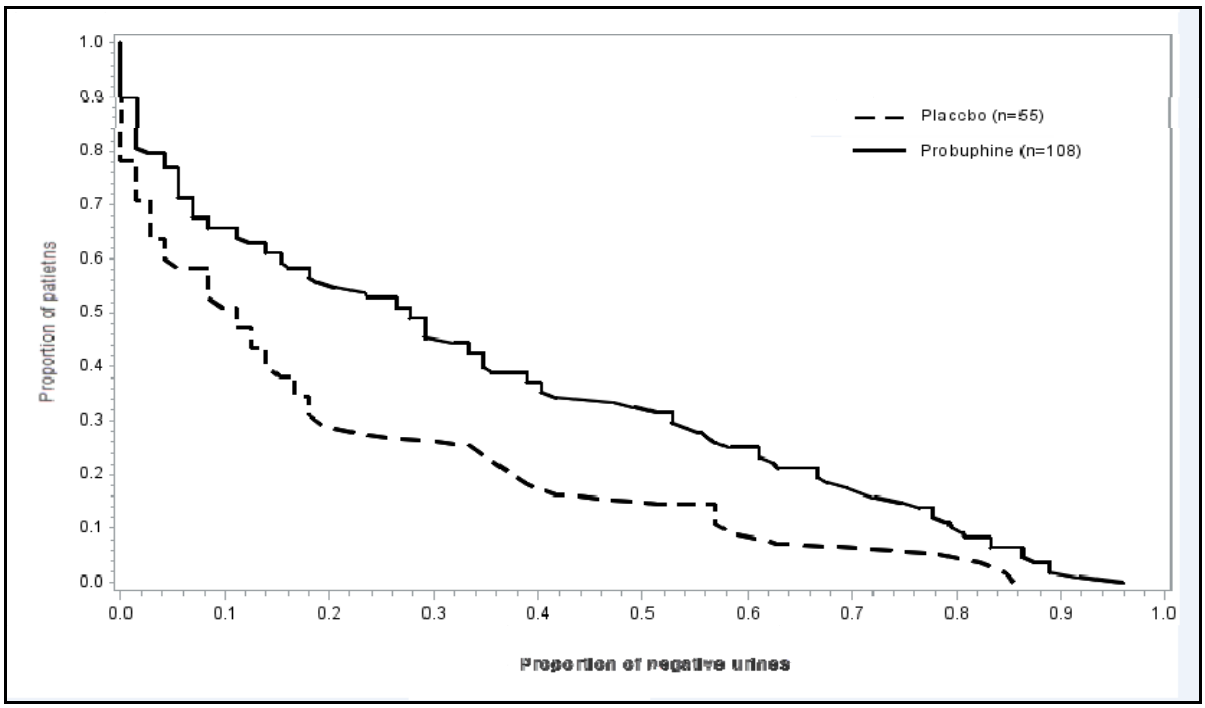
Applicant's data presentations, the curve would rise from the bottom left to the top right, instead of falling from the top left to the bottom right.

To provide a more intuitive presentation of the study results, Mr. Petullo graphed the data to illustrate the proportion of patients who submitted a particular percentage of negative tests *or better*. Thus, for any individual cutoff value (30%, 50%, 70%, etc., chosen for expedience and not because they are known to have any particular significance), there are fewer patients meeting each threshold. To facilitate comparisons at specific cutoffs, Mr. Petullo also provided tabulations of the proportion of patients meeting each cutoff.

For the Statistics Reviewer's analyses, the conventions used for urine sample and self-report data differed somewhat from the rules used by the Applicant'. Urine samples that the Applicant deemed non-missing and non-analyzable were included and considered positive for the purposes of the Statistics review. If a subject reported opioid use for the past two weeks, any negative urine tests were considered to be positive for those two weeks. Subjects were asked "have you used illicit opioids?" and "what was the duration of use?"

The figure below displays the CDF of percent negative urine samples for Weeks 1–24 with self-reported use incorporated generated by the Statistics Reviewer. The curves fall from 0% at the left to 100% at the right. For example, approximately 45% of the patients in the Probuphine group had at least 30% of urines samples negative for opioids. In comparison, approximately 27% of patients in the placebo group had at least 30% of urine samples negative for opioids.

PRO-805: CDF of the Percentage of Urine Samples Negative for Opioids in Weeks 1 – 24 with Incorporation of Subject Illicit Opioid Self-Report (ITT Population)



The CDF was statistically significantly different (p-value of 0.01) in PRO-805. In addition, there were more patients in the Probuphine arm that achieved at least 30%, 50%, or 80% negative urines. This information is provided in tabular format in the following table.

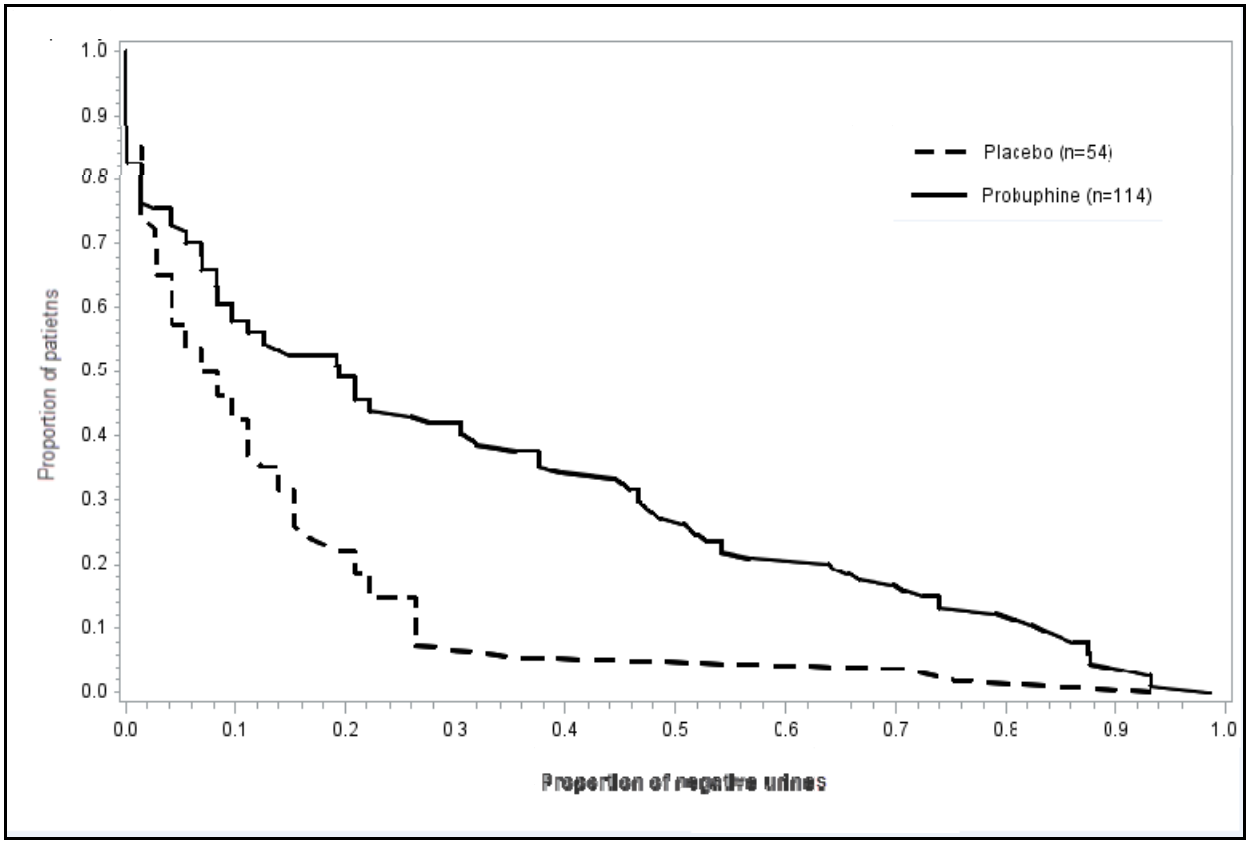
Study PRO-805: Percentage of negative urines, Weeks 1-24

Study	% Negative Urines	% of subjects	
		Probuphine	Placebo
PRO-805	≥ 30	45	27
	≥ 50	32	16
	≥ 75	15	7
	≥ 80	10	5
	≥ 85	6	2
	≥ 90	2	-
	≥ 95	1	-
	100	-	-

The cumulative distribution function and the tabular summary demonstrate that at each given level of percentage of opioid-negative urines, patients on Probuphine were more likely to submit opioid-negative urines. However, there were no patients in either treatment arm that achieved complete abstinence and few whose samples were opioid-negative more than half the time.

The CDF of percent negative urine samples for Weeks 1–24 for Study PRO-806 with self-reported use incorporated is shown in the figure below.

Study PRO-806: CDF of the Percentage of Urine Samples Negative for Opioids in Weeks 1 – 24 with Incorporation of Subject Illicit Opioid Self-Report (ITT Population)



The CDF was again statistically significantly different (p-value of <0.001) in PRO-806. In addition, there were again more patients in the Probuphine arm that achieved at least 30%, 50%, or 80% negative urines. This information is provided in tabular format in the following table.

Study PRO-806: Percentage of negative urines Weeks 1-24

Study	% Negative Urines	% of subjects	
		Probuphine	Placebo
PRO-806	≥ 30	42	7
	≥ 50	27	6
	≥ 75	13	4
	≥ 80	12	2
	≥ 85	9	2
	≥ 90	4	2
	≥ 95	1	-
	100	-	-

As observed in the PRO-805, the cumulative distribution function and the tabular summary demonstrate that at each given level of the percentage of negative urines, patients on Probuphine in Study PRO-806 were more likely to submit opioid-negative urines. The efficacy findings observed in PRO-805 were, in fact, replicated in Study PRO-806. However, again, there were no patients in either study that achieved complete abstinence and few whose samples were opioid-negative more than half the time.

In addition to the primary efficacy analyses for the time period of Weeks 1–24, the Applicant was encouraged to also look at the analyses of the endpoint allowing for a suitable "grace period." Recognizing that patients require some time for engagement in treatment, a grace period during which drug use is not counted in the assessment of response is permissible for the purposes of efficacy ascertainment. The Applicant chose two grace periods of four and eight weeks, reported a summary of the significance testing for each of the analyses for the pooled double-blind studies and the studies individually, and found statistically significant results across both timeframes. It is noteworthy, though, that no patients achieved complete abstinence when these grace periods of four and eight weeks were considered.

Mr. Petullo conducted analyses allowing for four months of grace (evaluating results based only on urine samples during Weeks 17-24), providing even more leniency with respect to allowing for engagement in treatment in order to assess for better outcomes. In Study PRO-805, there was one patient in the Probuphine arm who had no positive or missing urine samples in the final eight weeks, and in Study PRO-806, there were two. However, there was little indication that allowing four months for engagement in treatment produced a better picture of the results. This is in contrast to general clinical expectations that patients improve over time.

The review team also considered the possibility that three times a week urine testing may have been too burdensome. Patients who are successfully achieving abstinence from illicit drugs may well experience improvements in their social and occupational functioning that provide them with very legitimate reasons to miss study visits. To explore this, Mr. Petullo reanalyzed the data to determine the percentage of subjects who self-reported abstinence, and had negative results for all urine samples collected during each of the last 8 weeks of treatment. For example, if a subject provided a negative urine

sample during Visit 1 but missed Visits 2 and 3 of the same week, the subject was considered opioid-free for that week, unless the subject self-reported drug use. Results are shown below in Mr. Petullo's table.

Percent of Subjects With Self-Reported Abstinence and No Positive Samples, Weeks 17-24

	% Subjects	
Study	Placebo	Probuphine
PRO-805	0	6
PRO-806	2	4

Source: Statistics Reviewer

Although this analysis provides a more encouraging picture than the analyses which impute opioid-positive results to missing samples, it is nevertheless dismaying. Even using this most generous approach defining abstinence, and despite over 60% of Probuphine-treated subjects continuing to the end of the study with *ensured compliance with medication* due to the delivery system, only a very small fraction were able to attain abstinence after four months of grace and sustain it for two months of additional observation.

6.4.4 Use of Rescue Medication

As noted above, rescue medication could be provided at clinic visits when patients met protocol-specified criteria on the basis of withdrawal or craving scores. The table below illustrates the effect of different protocol-specified criteria for providing rescue. A markedly reduced proportion of patients received rescue in PRO-806, compared to PRO-805. Because the populations, procedures, and study medications were identical, it seems logical to conclude that this difference is attributable to the more stringent criteria for rescue applied in PRO-805.

Summary of Supplemental Buprenorphine Use (Intent-to-Treat Population)

Study	Treatment Group	Number(%) Subjects Requiring Supplemental SL	Number(%) Subjects Requiring Fifth Implant
PR0-805	Probuphine	67 (62.0)	22 (20.4)
	Placebo	50 (90.9)	32 (58.2)
PR0-806	Probuphine	45 (39.5)	25 (21.9)
	Placebo	36 (66.7)	21 (38.9)
	SL buprenorphine	7 (5.9)	Not allowed

SL = sublingual.

Notably, over half of the patients who qualified for a fifth implant in the two studies pooled continued to require rescue medication, although their mean days of rescue use per week and mean milligrams used per week declined.

The Applicant interpreted the use of rescue medication in these studies as an efficacy indicator, pointing out that more patients in the placebo-group required rescue medication than patients in the Probuphine group. However, this is certainly to be expected. All patients were physically dependent on opioids and those in the placebo arm had opioids abruptly discontinued. Indeed, the use of rescue for the placebo-treated patients was, to some extent, necessary for an ethical trial design, because patients seeking treatment for opioid addiction are almost always offered some pharmacologic treatment of their withdrawal symptoms. Not surprisingly, some placebo-treated patients required regular doses of rescue medication at each treatment visit, and quickly met criteria for "treatment failure." If the objective were to establish efficacy in treating symptoms of withdrawal, this finding would be encouraging. However, the objective was to demonstrate efficacy in maintenance treatment of opiate addiction, which implies that an effect on illicit drug use will be accomplished.

The Applicant also interpreted the frequency of rescue use in the clinical trials as support for their claim that Probuphine treatment would reduce the need for patients to have a supply of buprenorphine tablets or films in the home, which could translate to reductions in abuse, misuse, diversion, and accidental pediatric exposure. However, it must be stressed that the criteria for provision of rescue and the circumstances under which rescue doses were provided in the clinical trials bore very little relationship to the real-world scenario.

6.4.5 Graphic Depiction of Individual Patient Results

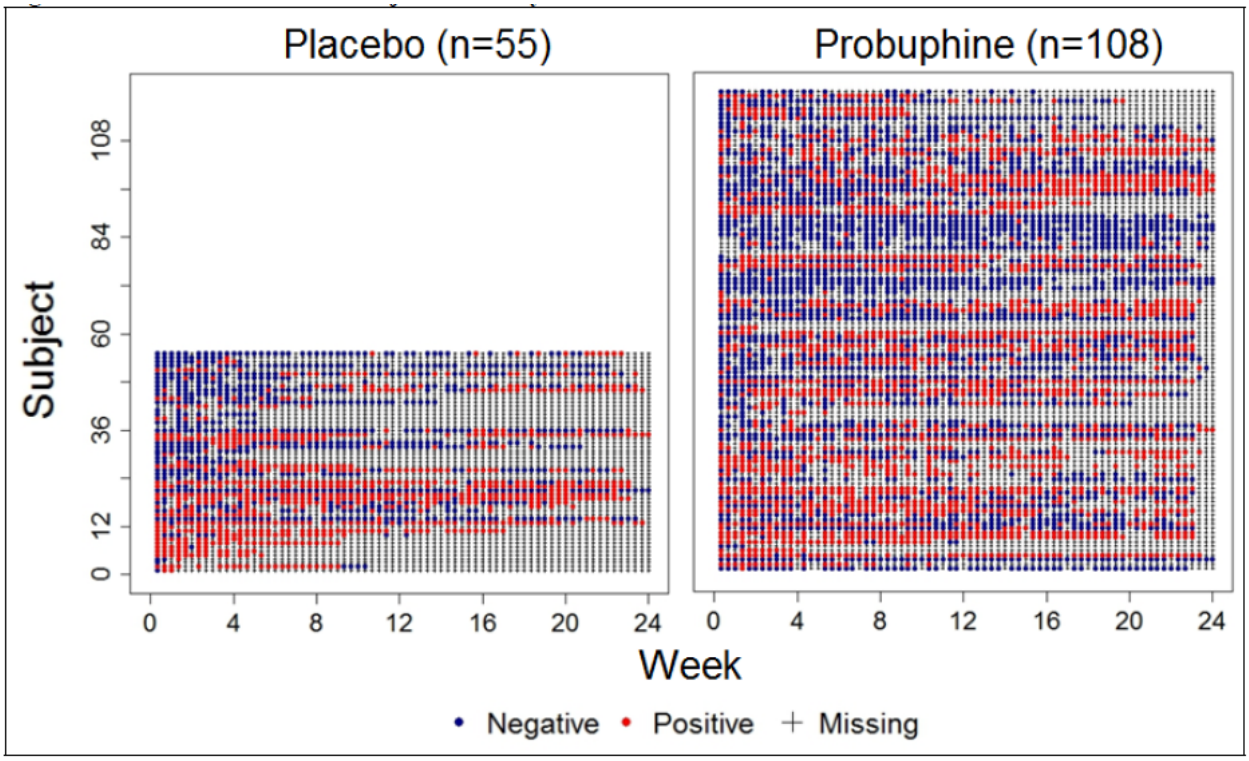
To allow an appreciation of the temporal sequence of patients' test results, Mr. Petullo prepared graphic depictions that show the results of each test for each patient. The overall percent of negative tests does not differentiate between, for example, a patient who is abstinent for half the study and then relapses to daily illicit drug use, a patient who continues to use illicit drugs daily for half the study and then stops completely, and a patient who uses intermittently, half the days throughout the study. All of these patients might have 50% of their tests negative. The graphic depictions distinguish among these patterns. They also distinguish between tests that were imputed as positive because they were missing, or because a patient self-reported drug use, and actual positive tests. Mr. Petullo also provided a graphic display of the use of rescue medication over time for each patient.

6.4.5.1 Urine Test Results

These subject-level analyses are shown below. In these presentations, each individual subject is represented along the y-axis. On the x-axis are the time points during which urine samples were collected. (In these studies, urine samples were collected three times per week). Blue dots are used to represent submission of opioid- negative urine samples

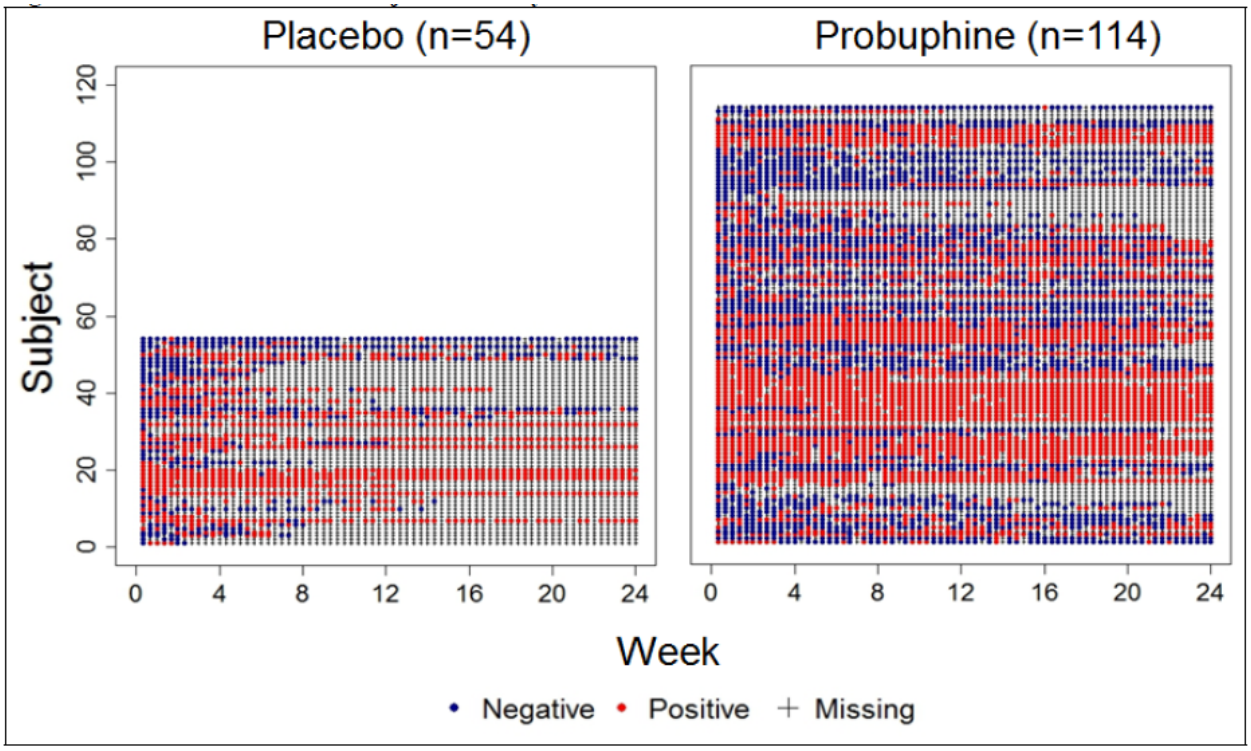
at any timepoint, while red dots are used to represent opioid-positive urine submissions. Ideally, a patient achieving treatment success would have many more blue data points than red data points, particularly along the right-hand side of the x-axis which represents longer periods of time on treatment. The data points that appear gray in these presentations are '+' symbols and denote missing urine data.

PRO-805 Subject-level Urine Sample Results



Source: Statistics Review

PRO-806 Subject-level Urine Sample Results



Source: Statistics Review

These figures illustrate a surprising result. The clinical expectation in a six-month period of addiction treatment, or a six-month study, is that the patient will probably either improve over time, or drop out of treatment. Completers are expected to be more successful than early dropouts, at least if the end of the study period is the time window of interest. In these studies, however, many patients remained in the study throughout the study period, consistently submitting opioid-positive urine samples over time. There do not seem to be many examples of patient gradually attaining greater periods of abstinence, or patients who have an early response but regrettably relapse.

In many addiction treatment studies, retention in treatment is one of the efficacy outcomes, based on an assumption that retention in treatment is a predictor of good outcome. These assumptions are derived from studies of patients on methadone treatment. These patients came to the clinic daily to receive their methadone dose, with visits potentially decreasing over time as the patients attained greater stability and time refraining from illicit drugs. Attendance at clinic also entailed participation in other aspects of addiction treatment apart from the pharmacological. It is reasonable to believe that there is some therapeutic benefit to coming to treatment visits.

Studies in patients on buprenorphine, too, also initially required daily supervised administration and regular clinic visits. Only since 2002 have patients treated with buprenorphine been able to receive treatment without very frequent clinic visits. However, studies buprenorphine-treated patients that evaluate retention in treatment also pertain to patients coming back to the study site, and participating in treatment visits that

may involve the provision of non-pharmacologic therapy, or may simply have the known therapeutic benefits of being in a treatment setting.

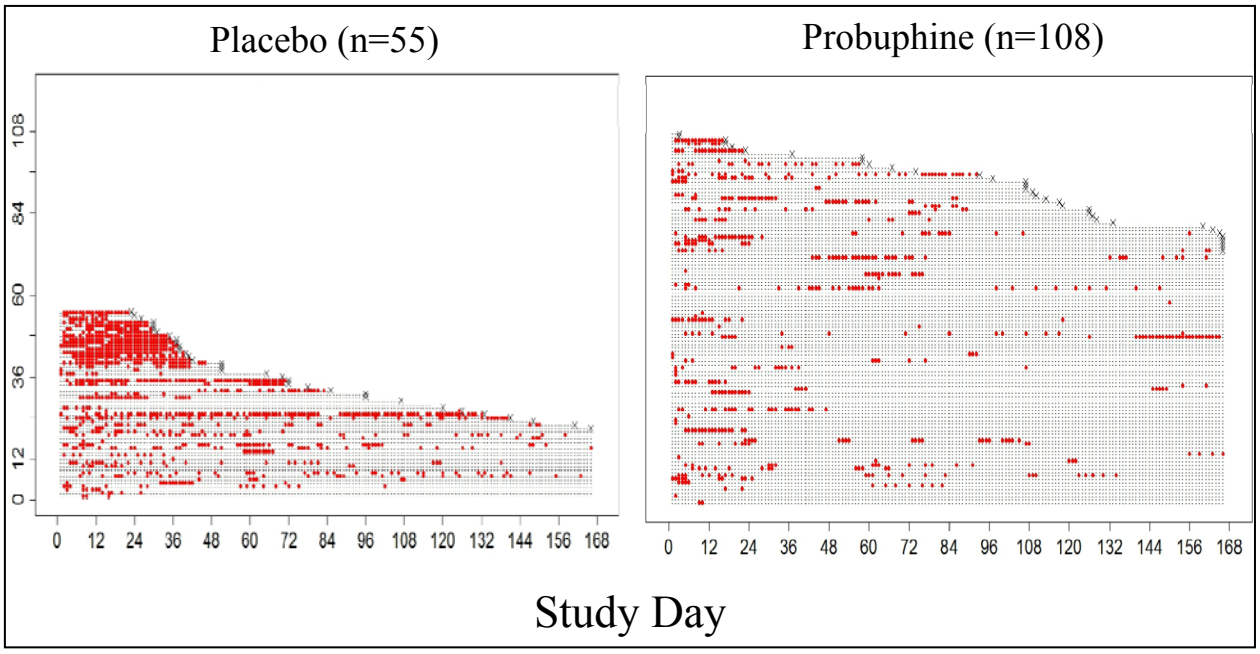
It is striking that patients in the Probuphine studies complied with regular clinic visits over six months, but the protocols may have provided some incentives for them to do so. For example, patients had the prospect of receiving rescue medication at any time and in fact continued to do so sporadically throughout the treatment period.

Conversely, in clinical practice, patients with Probuphine implants may not be retained in treatment. They will be *on treatment,* in the sense that they have circulating blood levels of buprenorphine, but they will not necessarily be *in treatment.* They may have no incentive at all to come to counseling visits or checkups with their treatment provider—and in fact, will be battling a disincentive in the form of charges for office visits. With no reinforcement, in the form of receiving their next monthly prescription, patients may not be seen at all. This called into question whether the benefits of "retention in treatment" would to accrue to these patients.

6.4.5.2 Use of Rescue Medication

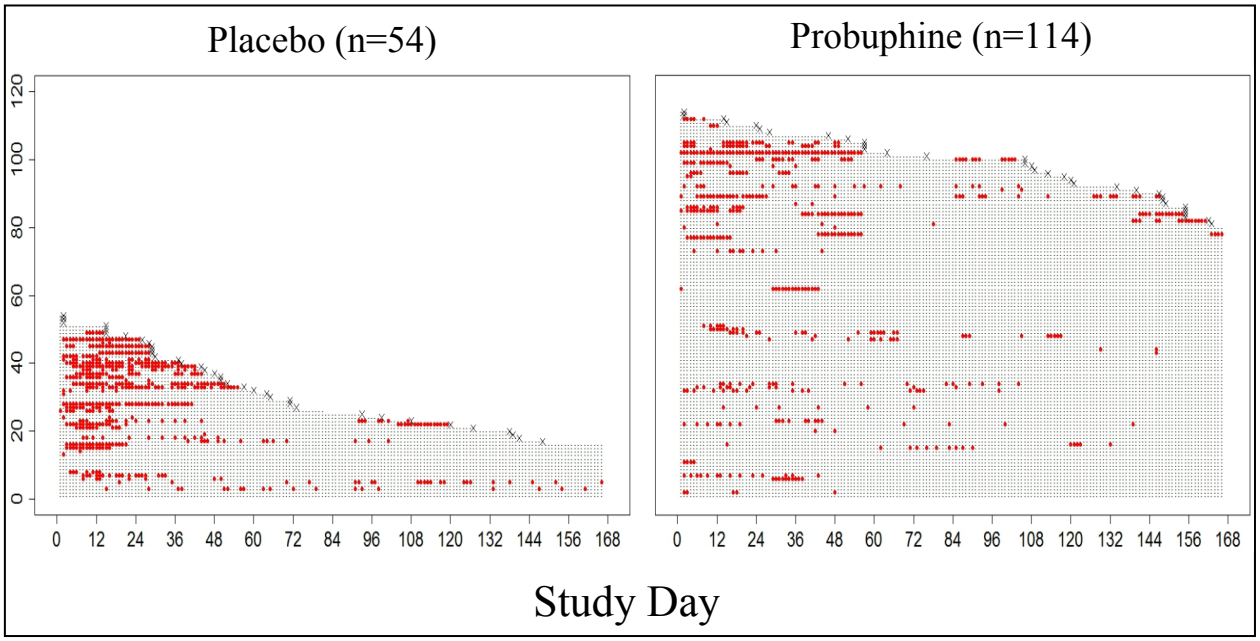
Graphic depictions of the use of rescue medication over time for individual patients are shown in Mr. Petullo's figures, below. On the y-axis, individual patients are represented. Patients are sorted by date of discontinuation. On the x-axis are the number of days in the trial. The red dots denote any use of sublingual buprenorphine on a particular day.

PRO-805: Individual Patient Use of Rescue Medication



Source: Statistics Reviewer-generated graphical displays

PRO-805: Individual Patient Use of Rescue Medication



Source: Statistics Reviewer-generated graphical displays

These figures show that rescue medication use was not limited to the early treatment period where dose titration would be expected to take place. It would not be surprising to see that a subset of patients might require rescue, indicating a need for a fifth implant, get the fifth implant, and require no more rescue. However, sporadic use of rescue medication continued throughout the study. As noted above, over half of patients who met criteria for up-titration with a fifth implant required rescue even after the additional implant was placed.

6.5 Discussion

In the end, the results of the controlled studies revealed that there were vanishingly few patients who attained a pattern of drug use that can convincingly be called successful treatment. Fully 64% of Probuphine-treated patients in Study 805 and 70% in Study 806 submitted opioid-negative urine samples on fewer than 30% of occasions in Weeks 17-24. Albeit, the numbers were even worse for placebo, and the imputation of positive samples for study dropouts inflates these numbers, but with a completion rate of ~65% for Probuphine-treated patients, data imputation is not the explanation. Parameters of drug use were similar between the Probuphine and open-label sublingual buprenorphine-treated arms in Study PRO-806—which is a disappointing result, because passive compliance formulations are expected to perform better than dosage forms that must be, but may not be, self-administered daily. If Probuphine overcomes the limitations of sublingual buprenorphine by ensuring compliance, one would expect it to be *better.* Our statute does not require superiority to a comparator, but a passive compliance delivery system makes the implicit claim that compliance, and therefore, efficacy, will be superior to daily dosing of the same drug. However, if the dose chosen is inadequate, then this

promise is not delivered upon. It is difficult to make comparisons between treatments for several reasons. First, patients volunteering for a clinical trial of a new, implantable formulation may, understandably, have been dismayed to an open-label arm in which they received a medication already available—potentially, one they'd already tried. Some patients may not have had access to buprenorphine treatment, or may have had difficulty paying for it, but, for others, there were certainly other less burdensome and less intrusive ways to received treatment with sublingual buprenorphine tablets than participation in a clinical trial that required multiple weekly visits. Furthermore, the protocol permitted only doses between 12 and 16 mg/day. Patients could not have dose escalations above 16 mg, and if the dose was decreased at any point, it could not be increased again. Therefore, the dosing was not individualized or titrated to effect, as patients would experience in the normal course of clinical practice. (It is acknowledged that a clinical trial does, customarily, provide dosing flexibility than "real-world" practice, but to the extent that we wish to know how Probuphine compares to "usual care," this becomes a relevant issue.) Although only 2% are described as discontinuing early to obtain other treatment for opioid dependence, the number of patients lost to follow-up was higher in this arm than in either of the other arms of the study. Nevertheless, overall retention was essentially identical to the Probuphine arm. Compliance with medication was not ensured (medication was given in 7-day take-home supplies).

It should be noted that in "real-world" practice, physicians are not blind to the results of their patients' toxicology screens, and can titrate the medication to effect or refer patients to more structured treatment if necessary. One of the selling points of a six-month implant is that patients need not be followed closely, monitored carefully, or have individually-titrated treatment. However, if these are not provided, then results may not be improved.

In trying to understand whether the results meet clinical expectations of "success," various sources in literature were reviewed. The treatment guidelines provided by SAMHSA provide a flow chart that includes dose increases when a patient continues to use illicit drugs; this would imply that ongoing drug use is not an expected, acceptable, and routine issue to be overlooked. But this may reflect an aspirational approach to treatment. Looking at how success and failure are defined in clinical trials, Weiss et al analyzed both a "good outcome" definition (abstinent during the final week of a 12-week treatment, and in at least two of the previous three weeks) and complete abstinence in a trial of buprenorphine vs. placebo in patients addicted to prescription opiates. (In this study, 34-39% were completely abstinent for the last four weeks.) On the opposite side of the coin, Fiellin et al conducted two 24-week studies comparing buprenorphine treatment under different conditions of ancillary behavioral therapy. In these protocols, dose increases were allowed for patients with "evidence of ongoing (for 3 consecutive weeks) illicit opioid use," and the protocols stipulated that "patients with unremitting illicit drug use (3 consecutive weeks of urine specimens positive for opioids after the buprenorphine dose had been increased to 24 mg) met criteria for protective transfer." Protective transfer refers to discontinuation from the protocol and referral to more structured and intensive treatment. This indicates that ongoing use of illicit drugs was not an expected outcome of treatment. As 30% of participants met criteria for protective transfer in one study (2013),

and 11% in the other (2006), this suggests that 70-89% of participants did *not* have ongoing illicit opioid use. Mitchell, et al reported outcomes for 300 patients entering community-based buprenorphine treatment and followed over 6 months. At six months, 56% in a standard outpatient treatment group and 49% in an intensive outpatient treatment group had opioid-positive urine tests. However, days of heroin use in the past 30 days declined from 22 to 3-4. Thus, even in a community treatment setting (not a clinical trial), while abstinence is less common than one would hope, it is certainly higher than in the Probuphine study and patients who continue to use on 3-4 days/month would be likely to submit negative urine samples about ¾ of the time.

Comparisons to the historical data on the efficacy of buprenorphine are very challenging. Those studies were performed under very different conditions, and the data are not available to subject to analysis of the cumulative distribution functions. One of the trials compared 8 mg sublingual solution (considered roughly equivalent to a 12 mg tablet dose) to two doses of methadone, 20 mg (likely sub-therapeutic) and 60 mg (considered to be the low end of the therapeutic range). The other study compared sublingual buprenorphine solution at 1 mg, 4 mg, 8 mg, and 16 mg (considered roughly equivalent to <2 mg, 6 mg, 12 mg, and 24 mg as tablet doses). Patients were new entrants to treatment, and all were heroin users. Studies involved titration, 4 months of maintenance, and then either taper or open-label follow-on. The results of these studies, in terms of measures of retention and group mean percent opioid-negative urine samples, were not even as encouraging as the results in the open-label arm of Study PRO-806. The populations may have differed (100% heroin-dependent vs. ~60%), and the registration studies for Subutex required daily clinic visits for supervised administration, which is a burdensome feature. The studies were nevertheless accepted as substantial evidence of efficacy; however, it must be noted that the intention was that the medication would be titrated to effect and that patients who did not cease illicit drug use might need higher doses, more structured treatment, or different treatment altogether.

6.5.2 Dose-Response Issues

As noted above, the clinical experience with buprenorphine as it is currently used yields a higher expectation of efficacy. It is not very surprising that the efficacy results for Probuphine are discouraging compared to the expected efficacy of buprenorphine, based on various sources of information. To begin with, the plasma level of buprenorphine in patients treated with Probuphine is half the trough level associated with a 16 mg/day dose of buprenorphine sublingual tablets, which is the target dose recommended in labeling of Subutex and Suboxone. Based on AUC, the level is only 31% that of 16 mg/day dosing. One might wonder why this dose was sufficient to hold patients in treatment, and why there was as little use of supplemental buprenorphine as was observed (about half the patients). This may be explained in two ways.

First, the dose of buprenorphine necessary to allay opioid withdrawal symptoms is very low. Before Suboxone and Subutex were approved, Buprenex (parenteral buprenorphine) was fairly widely used off-label for treatment of withdrawal, at doses of 0.1-0.2 mg i.m. Other data (discussed in Section 3, above) suggest that there is a substantial difference between the level of mu opioid receptor occupancy associated with blockade of

exogenous opioids and that associated with withdrawal, confirming that the pharmacokinetic/pharmacodynamic relationships are different for these two effects of buprenorphine.

On the other hand, one reason patients drop out of treatment is that they would like to take a "vacation" from buprenorphine in order to experience the effects of their drug of choice. Buprenorphine, given at the doses recommended as the target dose in labeling, is intended to block the effects of exogenous opioids. The dose needed to accomplish this effect is much higher than the dose needed to treat withdrawal. It is widely accepted that effectively-blocked patients may "test the blockade" but do not typically engage in regular illicit drug use, because it is "a waste of money." The dose of buprenorphine provided by Probuphine appears to have allowed a substantial fraction of the patients to continue using illicit opioids occasionally, or even regularly, without needing to discontinue treatment.

6.5.3 Clinical Interpretation of Results

As to the reason for the relatively low use of rescue buprenorphine, it must be noted that the rescue doses were available only at clinic visits, and only when patients met specific criteria, which were related to withdrawal and "craving." Patients may well have been without measurable symptoms on these instruments, but control of subjective symptoms of withdrawal and "craving" is of uncertain relevance if it does not translate to drug use behavior. One can assume, additionally, that clinicians monitoring their patients' urine toxicology results (unlike the site investigators, who were blind to the results), would have concluded that the great majority of patients needed rescue medication due to ongoing illicit substance use.

To understand how is this different from other conditions, in which we sometimes accept any difference between the drug-treated patients and the placebo-treated patients as beneficial, it must be reiterated that opioid addiction is a complex of behavioral experiences, in which patients are compelled to use opioids despite ongoing harm, experience preoccupation with thinking about obtaining, using, and recovering from opioid use, and give priority to opioid use over other life activities, to the detriment of their health, psychosocial wellbeing, and occupational functioning. In an analgesic trial, the problem is pain. The symptom that is being measured is pain. If pain is reduced from baseline by the test drug more than the control drug, we can conclude that pain—the problem—is being treated. Patients using the medication and their physicians can readily ascertain whether the problem is being treated well enough to continue on that medication or not. In the Probuphine studies, we were looking at a surrogate endpoint of uncertain predictive value. The data provide little insight into what level of ongoing drug use could be used to conclude that the patient's opioid addiction was responding or not responding to treatment.

6.5.4 Enrichment Strategy

It should also be noted that these studies employed an enrichment design. Only patients who could tolerate buprenorphine, and who could be stabilized on a dose of 12 mg-16 mg

of buprenorphine within about 10 – 16 days could be enrolled. If anything, this design should give a more optimistic picture of the product's efficacy (and safety) than use in a general population. A significant number of patients (e.g., 84 of 115 screen failures in PRO-805) were screened out based on this criterion, suggesting that many patients will not meet the criteria that will be described in labeling. In the clinical trials, a significant number of waivers were granted to allow patients not meeting the run-in dose criterion to enroll (e.g., 72 of 83 waivers granted in PRO-805), but it is not clear whether these patients required additional time, less time, lower doses, or higher doses than the target.

6.6 Conclusion

In summary, despite an enrichment strategy which enrolled patients considered responsive to buprenorphine, only a very small minority of patients treated with Probuphine at the recommended dose seem to have accomplished substantial improvements in their drug-use behavior, even over six to twelve months of treatment. Taken together, concerns about the clinical significance of the primary analysis, pharmacological reasons to doubt the dose would be effective in blocking exogenous opioids, and the expectation that in the "real world" clinical setting, almost every patient will require ongoing sublingual buprenorphine to supplement Probuphine treatment, led the review team to the conclusion that the benefits of Probuphine, at the dose tested, did not outweigh the risks for the population studied.

Advisory Committee Meeting

In order to gain a better understanding of the risk/benefit balance for Probuphine, a meeting of the Psychiatric Drugs Advisory Committee was held on 3/22/13 to discuss the Probuphine application. Although the majority of the committee voted that efficacy had been demonstrated, that safety had been adequately characterized, and that the risk/benefit ratio favored approval, the comments during the discussion and the breakdown of votes revealed considerable ambivalence about the application.

Many participants, even some who voted that efficacy had been demonstrated, expressed that their vote reflected the fact that, on the primary endpoint, the drug had out-performed placebo. Several did note concerns about the adequacy of the dose, and five voted that efficacy had not been demonstrated. Panel members noted difficulty reconciling Applicant's claim that the steady-state blood levels were maintained in an efficacious range with the pattern of urine toxicology results, asking "How can I make the claims of robust efficacy jive with the very disappointing results in terms of negative urines?" and "I'm not sure that we're doing anyone a service if we put something on the market that's not the right dose, that doesn't actually optimally achieve what we're trying to accomplish..," and "if there's tons of positives at the 24th week, did that medicine do the right thing, or … what's the purpose of that drug?" One panelist noted that, if he treats a patient with buprenorphine "And if they've got a few months of dirty urine, I'm going to say treatment failed" and refer the patient for other treatment.

Although addiction medicine providers on the panel observed that it was not a requirement or expectation for a patient in treatment to have "totally clean urines," one provider noted that "when we have people who give us urines that are positive for illicit or unauthorized use, it's a signal to us that we need to reevaluate the patient and make some changes in whatever therapy that we're providing." He expressed some concern about the use of supplemental buprenorphine as an indication that the dose was inadequate, and concern that providers would be "stuck with" a dose that would leave providers with "difficulty meeting our patients' needs." Another addiction medicine expert noted that the REMS would need to address the use of supplemental sublingual buprenorphine and that physicians would need to be educated to minimize its use.

An expert appearing on behalf of the Applicant noted that "In thinking about the individuals who were stabilized on 12 to 16 milligrams and then transferred over to the rods, I would look at them as very early responders to the treatment. So they're in the phase 1 and 2, at most, in the six months of study. Her further comments seemed to indicate that patients need to be titrated to the dose of medication that is necessary to help them discontinue illicit drug use, which may be higher than 12-16 mg, and certainly higher than the dose provided by Probuphine (which is 1/3 the AUC of 16 mg/day), noting, "people with mild to moderate disease, being those that you want to capture in the 12 to 16 range. I think that's where we want to induce people, and we want to increase them…until we get them at a place where we can reliably support their desire not to use heroin or prescription drugs while they're on this medication."

One addiction medicine specialist noted, "One of the populations that it has been suggested by several people, and I think it's an appropriate suggestion, is for the patient who's already stable. Half of my stable patients are on less than 12 milligrams. Most of them are on 4 or 8…I frankly think that this is an extremely important product concept, to be able to give a patient six months of medication that will keep them stable and that we would have limited oral or sublingual supplementation on that would decrease that issue with diversion…I've got patients who have been on buprenorphine sublingually now for five to eight years. None of them are on 16. My new patients are generally on 16, but they back off within six months, nine months to a lower dose.

Several participants in the open public hearing were investigators in the clinical trials or individuals with expertise in addiction treatment. Their comments reflected an expectation that the product would be efficacious enough to bring patients' addiction into remission, emphasizing that the benefit of the product was the six month duration of action, so that treatment would be ensured over a time sufficient to accomplish this goal. The comments did not address the clinical significance of the results in patients who continued to use illicit drugs persistently throughout the six months. One site investigator felt that the product would be appropriate for patients who were already engaged in treatment, stabilized, and no longer using illicit drugs for a year or more. Several commenters noted that Probuphine would facilitate treatment in patients who could not come to office visits—citing the possibility of telemedicine in rural communities, benefits for patients who travel, obviation of transportation problems. These commenters saw an advantage in the fact that Probuphine-treated patients would need to be seen only

infrequently. Conversely, other commenters emphasized the importance of treatment visits—noting that a medication that did not have to be taken daily would (paraphrasing) "help patients take steps toward focusing not on taking medication but on recovery; they can focus on remaining in treatment," and "allowing patients to dedicate time and attention to the psychosocial aspects of treatment," and that medication and therapy are both necessary, with medication helping the patient to refrain from illicit drugs, "the longer time away from the drug of choice, the more available the patient is for treatment."

Regarding safety, the discussion focused primarily on issues related to the insertion and removal procedures. The obstetrician/gynecology experts, Drs. Espy and Hewitt, emphasized that removal is the more difficult of the two procedures, but that complications of removal are often attributable to errors in insertion. They observed that the "U-technique" that is to be used in Probuphine removal is not the procedure that was used to remove Norplant; therefore, there is little experience with this procedure even among Norplant-experienced providers. The Applicant's expert on the insertion and removal procedures, Dr. Chavoustie, explained that the Probuphine implant is less "forgiving" (understood to mean more friable) than the Norplant implants, and therefore the alternate technique facilitates removal. Dr. Hewitt noted that "While I do think it's an easier skill for people to acquire that are comfortable doing surgical interventions, I feel really strongly that with the correct training that this is something that you can teach any provider to know how to do…..It is really important that the training be adequate and appropriate." Several commenters noted that "high volume" is important in developing and maintaining expertise in any procedure, and noted that certification should be reviewed if providers do not do the procedures regularly. The OB/Gyn experts also observed that providers should be required to have the ability to refer to someone who can do removals of deep implants, which, it was noted, is a specialized skill typically provided at a limited number of facilities.

Appendix C Clinical Stability Checklist

PRO-814 Checklist for Clinical Stability

A potential study subject must be considered clinically stable by their treating healthcare provider and confirmed by the following attestation:

Patient Name _____ Treating Physician Name_____

Treating Physician Address_____

_____

I consider this patient clinically stable based on the following (please check all that apply)

a. Patient has not reported any illicit opioid drug use in the past 90 days_____

b. Patient has a stable living environment_____

c. Patient participates in a structured activity/job that contributes to the community_____

d. Patient has not reported significant withdrawal symptoms in the past 90 days_____

e. Patient has consistently participated in recommended cognitive behavioral therapy/peer support program_____

f. Patient has been consistently compliant with clinic visit requirements_____

g. Patient has reported low to no desire/need to use illicit opioids in the past 90 days_____

h. No episodes of hospitalizations (addiction or mental health issues), emergency room visits, or crisis interventions in the previous 90 days

i. Please describe any other indicators of clinical stability that you have observed_____

_____

_____

Treating Physician Signature_____

Source: PRO-814 Manual of Procedures

Appendix D Common Adverse Events in buprenorphine studies from approved labeling

ADVERSE REACTIONS

In a comparative study, adverse event profiles were similar for subjects treated with 16 mg buprenorphine and naloxone sublingual tablets or 16 mg buprenorphine HCl sublingual tablets. The following adverse events were reported to occur by at least 5% of patients in a 4-week study (Table 3).

Adverse Events (≥ 5%) by Body System and Treatment Group in a 4-week Study

Body System /Adverse Event (COSTART Terminology)	N (%) Buprenorphine and Naloxone Sublingual Tablets 16 mg/day N=107	N (%) Buprenorphine HCl Sublingual Tablets 16 mg/day N=103	N (%) Placebo N=107
Body As A Whole			
Asthenia	7 (6.5%)	5 (4.9%)	7 (6.5%)
Chills	8 (7.5%)	8 (7.8%)	8 (7.5%)
Headache	39 (36.4%)	30 (29.1%)	24 (22.4%)
Infection	6 (5.6%)	12 (11.7%)	7 (6.5%)
Pain	24 (22.4%)	19 (18.4%)	20 (18.7%)
Pain Abdomen	12 (11.2%)	12 (11.7%)	7 (6.5%)
Pain Back	4 (3.7%)	8 (7.8%)	12 (11.2%)
Withdrawal Syndrome	27 (25.2%)	19 (18.4%)	40 (37.4%)
Cardiovascular System			
Vasodilation	10 (9.3%)	4 (3.9%)	7 (6.5%)
Digestive System			
Constipation	13 (12.1%)	8 (7.8%)	3 (2.8%)
Diarrhea	4 (3.7%)	5 (4.9%)	16 (15%)
Nausea	16 (15%)	14 (13.6%)	12 (11.2%)
Vomiting	8 (7.5%)	8 (7.8%)	5 (4.7%)
Nervous System			
Insomnia	15 (14%)	22 (21.4%)	17 (15.9%)
Respiratory System			
Rhinitis	5 (4.7%)	10 (9.7%)	14 (13.1%)
Skin And Appendages			
Sweating	15 (14%)	13 (12.6%)	11 (10.3%)

The adverse event profile of buprenorphine was also characterized in the dose-controlled study of buprenorphine solution, over a range of doses in four months of treatment. Table 4 shows adverse events reported by at least 5% of subjects in any dose group in the dose-controlled study.

Adverse Events (≥ 5%) by Body System and Treatment Group in a 16-week Study

Body System /Adverse Event (COSTART Terminology)	Buprenorphine Dose*				
	Very Low* (N=184)	Low* (N=180)	Moderate* (N=186)	High* (N=181)	Total*(N=731)
	N (%)	N (%)	N (%)	N (%)	N (%)
Body as a Whole					
Abscess	9 (5%)	2 (1%)	3 (2%)	2 (1%)	16 (2%)
Asthenia	26 (14%)	28 (16%)	26 (14%)	24 (13%)	104 (14%)
Chills	11 (6%)	12 (7%)	9 (5%)	10 (6%)	42 (6%)
Fever	7 (4%)	2 (1%)	2 (1%)	10 (6%)	21 (3%)
Flu Syndrome	4 (2%)	13 (7%)	19 (10%)	8 (4%)	44 (6%)
Headache	51 (28%)	62 (34%)	54 (29%)	53 (29%)	220 (30%)
Infection	32 (17%)	39 (22%)	38 (20%)	40 (22%)	149 (20%)
Injury Accidental	5 (3%)	10 (6%)	5 (3%)	5 (3%)	25 (3%)
Pain	47 (26%)	37 (21%)	49 (26%)	44 (24%)	177 (24%)
Pain Back	18 (10%)	29 (16%)	28 (15%)	27 (15%)	102 (14%)
Withdrawal Syndrome	45 (24%)	40 (22%)	41 (22%)	36 (20%)	162 (22%)
Digestive System					
Constipation	10 (5%)	23 (13%)	23 (12%)	26 (14%)	82 (11%)
Diarrhea	19 (10%)	8 (4%)	9 (5%)	4 (2%)	40 (5%)
Dyspepsia	6 (3%)	10 (6%)	4 (2%)	4 (2%)	24 (3%)
Nausea	12 (7%)	22 (12%)	23 (12%)	18 (10%)	75 (10%)
Vomiting	8 (4%)	6 (3%)	10 (5%)	14 (8%)	38 (5%)
Nervous System					
Anxiety	22 (12%)	24 (13%)	20 (11%)	25 (14%)	91 (12%)
Depression	24 (13%)	16 (9%)	25 (13%)	18 (10%)	83 (11%)
Dizziness	4 (2%)	9 (5%)	7 (4%)	11 (6%)	31 (4%)
Insomnia	42 (23%)	50 (28%)	43 (23%)	51 (28%)	186 (25%)
Nervousness	12 (7%)	11 (6%)	10 (5%)	13 (7%)	46 (6%)
Somnolence	5 (3%)	13 (7%)	9 (5%)	11 (6%)	38 (5%)
Respiratory System					

Cough Increase	5 (3%)	11 (6%)	6 (3%)	4 (2%)	26 (4%)
Pharyngitis	6 (3%)	7 (4%)	6 (3%)	9 (5%)	28 (4%)
Rhinitis	27 (15%)	16 (9%)	15 (8%)	21 (12%)	79 (11%)
Skin and Appendages					
Sweat	23 (13%)	21 (12%)	20 (11%)	23 (13%)	87 (12%)
Special Senses					
Runny Eyes	13 (7%)	9 (5%)	6 (3%)	6 (3%)	34 (5%)

*Sublingual solution. Doses in this table cannot necessarily be delivered in tablet form, but for comparison purposes: “Very low” dose (1 mg solution) would be less than a tablet dose of 2 mg; “Low” dose (4 mg solution) approximates a 6 mg tablet dose; “Moderate” dose (8 mg solution) approximates a 12 mg tablet dose; “High” dose (16 mg solution) approximates a 24 mg tablet dose.

Department of Health and Human Services
Public Health Service
Food and Drug Administration
Center for Drug Evaluation and Research
Office of Surveillance and Epidemiology
Office of Medication Error Prevention and Risk Management

Date: December 15, 2015

To: Members of the Psychopharmacologic Drugs Advisory Committee

From: Division of Risk Management
Office of Medication Error Prevention and Risk Management
Office of Surveillance and Epidemiology (OSE)

Subject: Risk Management Considerations

Product: Probuphine (buprenorphine HCl) implants for subdermal administration (NDA 204442)

1 INTRODUCTION

This memorandum provides an analysis of the risk mitigation strategies necessary to address the risks of migration, protrusion, expulsion and nerve damage associated with the improper insertion and removal of Probuphine, as well as, the risks of accidental overdose, misuse and abuse if an implant comes out or protrudes from the skin.

2 BACKGROUND

2.1 PRODUCT INFORMATION

Probuphine is a schedule III, buprenorphine-containing subdermal implant. The Applicant is seeking approval of Probuphine for the maintenance treatment of opioid dependence, to be used as part of a complete treatment program, including counseling and psychosocial support.

Probuphine is available as a 26 mm x 2.5 mm rod-shaped implant and contains 80 mg buprenorphine HCl. Four rods are to be implanted subdermally at the inner side of the upper arm (about 8-10 cm above the medial epicondyle of the humerus). The implant provides sustained delivery of buprenorphine for up to six months. Once removed, new implants can be inserted in the opposite arm if continued therapy with Probuphine is warranted. Probuphine was developed as an alternative for practitioners and patients in the office based setting utilizing an abuse deterrent formulation. Probuphine is intended for use in patients who are opioid-tolerant and stabilized on daily doses of 8 mg or less of sublingual buprenorphine.

2.2 DRUG ADDICTION TREATMENT ACT OF 2000 (DATA 2000)[1]

The Drug Addiction Treatment Act of 2000 (DATA 2000) allows qualified physicians to obtain a waiver from the registration requirements of the Controlled Substances Act (CSA) to prescribe and dispense opioid medications in Schedule III, IV, and V for the treatment of opioid addiction provided such medications are approved by FDA for that indication. To qualify for a waiver under DATA 2000, physicians must hold a current state medical license, a valid registration number with the Drug Enforcement Agency (DEA), and meet any one or more of the following criteria and provide supporting document for all that apply:

- The physician holds a subspecialty board certification in addiction psychiatry from the American Board of Medical Specialties.
- The physician holds an addiction certification from the American Society of Addiction Medicine.
- The physician holds a subspecialty board certification in addiction medicine from the American Osteopathic Association.
- The physician has completed not less than eight hours of training with respect to the treatment and management of opioid-addicted patients. This training can be provided through classroom situations, seminars at professional society meetings, electronic communications, or otherwise. The training must be sponsored by one of five organizations authorized in the DATA 2000 legislation to sponsor such training, or by any other organization that the Secretary of the Department of Health and Human Services (the Secretary) determines to be appropriate.
- The physician has participated as an investigator in one or more clinical trials leading to the approval of a narcotic drug in Schedule III, IV, or V for maintenance or detoxification treatment, as demonstrated by a statement submitted to the Secretary by the sponsor of such approved drug.
- The physician has other training or experience, considered by the state medical licensing board (of the state in which the physician will provide maintenance or detoxification treatment) to demonstrate the ability of the physician to treat and manage opioid-addicted patients.
- The physician has other training or experience the Secretary considers demonstrates the ability of the physician to treat and manage opioid-addicted patients.

To obtain a waiver a qualified physician must notify the Center for Substance Abuse Treatment (CSAT), a component of the Substance Abuse and Mental Health Services Administration (SAMHSA), of their intent to begin dispensing or prescribing this treatment and contain their qualifications required to do so. The physician must also attest that they will refer addiction treatment patients for appropriate counseling and other non-pharmacologic therapies and will have no more than 30 addiction treatment patients

[1] Substance Abuse and Mental Health Services Administration. Buprenorphine – Drug Addiction Treatment Act of 2000. Available at: http://buprenorphine.samhsa.gov/titlexxxv.html Accessed February 20, 2013.

under their care at any one time unless, at least one year from the date the physician provided initial notification, a second notification is submitted to the Secretary stating the need and intent to treat up to 100 patients.

2.3 RISK EVALUATION AND MITIGATION STRATEGIES[2]

Section 505-1 of the Food, Drug, and Cosmetic Act (FDCA) authorizes the FDA to require pharmaceutical applicants to develop and comply with a risk evaluation and mitigation strategy (REMS) for a drug if FDA determines that a REMS is necessary to ensure that the benefits of the drug outweigh the risks. A REMS is a required risk management plan that uses risk minimization strategies beyond the professional labeling. The elements of a REMS can include: a Medication Guide or patient package insert (PPI), a communication plan to healthcare providers, elements to assure safe use, and an implementation system. FDAAA also requires that all REMS approved for drugs or biologics under New Drug Applications (NDA) and Biologics License Applications (BLA) have a timetable for submission of assessments of the REMS. These assessments are prepared by the sponsor and reviewed by FDA.

Elements to assure safe use (ETASU) can include one or more of the following requirements:

- Healthcare providers who prescribe the drug have particular training or experience or special certifications
- Pharmacies, practitioners, or healthcare settings that dispense the drug are specially certified
- The drug may be dispensed only in certain healthcare settings
- The drug may be dispensed to patients with evidence of safe-use conditions
- Each patient must be subject to monitoring
- Patients must be enrolled in a registry

Because ETASU can impose significant burdens on the healthcare system and reduce patient access to treatment, ETASU are required only if FDA determines that the product could be approved only if, or would be withdrawn unless, ETASU are required to mitigate a specific serious risk listed in the labeling. Accordingly, the statute [FDCA 505-1(f)(2)] specifies that ETASU:

- Must be commensurate with specific serious risk(s) listed in the labeling.
- Cannot be unduly burdensome on patient access to the drug.

To minimize the burden on the healthcare delivery system, must, to the extent practicable, conform with REMS elements for other drugs with similar serious risks and be designed for compatibility with established distribution, procurement, and dispensing systems for drugs

---

[2] FDA Draft Guidance for Industry – *Format and Content of Proposed Risk Evaluation and Mitigation Strategies (REMS), REMS Assessments, and Proposed REMS Modifications*, dated September 2009. Available at: http://www fda.gov/downloads/Drugs/Guidances/UCM184128.pdf.

2.4 APPROVED REMS FOR BUPRENORPHINE PRODUCTS FOR THE TREATMENT OF OPIOID DEPENDENCE

Buprenorphine products were the first narcotic drugs available for the treatment of opioid dependence in an office-based treatment program under DATA-2000. Currently, buprenorphine HCl for the treatment of opioid dependence in an office-based treatment program under DATA-2000 is available in sublingual (SL) tablets. In combination with naloxone HCl, an opioid antagonist, the dosage forms include SL tablets, SL film and buccal film. [3]

All buprenorphine products approved for the treatment of opioid dependence in an office-based treatment program under DATA-2000 are approved with a REMS to ensure the benefits of the drug outweigh the risks.[4] In particular, the Agency determined that these products could only be approved if ETASU were required as part of a REMS to mitigate the risks of (1) exposure in persons for whom it was not prescribed, including accidental exposure in children, and (2) risks of abuse and misuse, listed in the labeling. The ETASU informs patients of the serious risks associated with buprenorphine products and the appropriate conditions of safe use and storage of buprenorphine products. The ETASU also ensures adequate clinical monitoring of patients by healthcare providers.

The goals of the REMS for Buprenorphine products are to:

- Mitigate the risks of accidental overdose, misuse and abuse
- Inform patients, prescribers and pharmacists of the serious risks associated with buprenorphine products.

The elements of the REMS include a Medication Guide, ETASU that include documentation of safe use conditions and ongoing monitoring requirements, and an implementation system. The REMS does not link prescribing or dispensing to documentation of safe use conditions and monitoring elements (e.g., is not a restricted distribution program).

3 SERIOUS SAFETY CONCERNS FOR PROBUPHINE

Due to the formulation of Probuphine, it is associated with complications related to improper technique associated with the insertion and removal procedure. Complications related to the procedure include those common to minor surgeries, such as pain, infection, bleeding and scarring. The potential risks serious enough to warrant mitigation beyond labeling include migration, protrusion, expulsion and nerve damage. In the clinical trials, at least six patients experienced expulsions or extrusions of an implant. If an implant is

[3] Buprenorphine for the management on pain is available as an extended-release transdermal patch (Butrans®) and a solution for injection. Buprenorphine/naloxone combination product for the management of pain is available in a buccal film (Belbuca®).

[4] There are 4 approved REMS for buprenorphine products: Suboxone SL tablet, Suboxone SL film, Subutex SL tablet and Buprenorphine for the Treatment of Opioid Dependence (BTOD).

expelled, there is the potential for accidental exposure and infections. If a Probuphine implant migrates, there is the potential that removal would be more complicated including necessitating use of either magnetic resonance imaging (MRI) or an ultrasound to locate the implant. Additional significant safety concerns include injuries related to damage of the ulnar or medial cutaneous nerve, which has led to permanent disability in implantable contraception devices with insertion and removal procedures similar to Probuphine. Unlike implantable contraception devices, Probuphine may be inserted or removed by practitioners who do not have extensive surgical backgrounds. And compared to contraceptive implants, higher incidences of bleeding (10.9%), complicated removals (3.2%), and implant site infection (4.0%) were noted in the Probuphine trials.[5] Adequate training on the insertion and removal procedure is essential to the minimization of post-surgery adverse effects.

The implant formulation has the potential of making accidental exposure, misuse and abuse more difficult because the buprenorphine is implanted subdermally. The decreased likelihood of accidental exposure is a potential advantage of its formulation. However, while the formulation may reduce the risk for accidental overdose, misuse, and abuse, these risks are not eliminated. Should an implant come out or protrude from the skin, the risks are still present. In the clinical trials, at least eight patients experienced expulsions or extrusions of an implant; therefore it is important to address this risk.[6]

If Probuphine is approved, a risk mitigation strategy (beyond professional labeling) will be required to address (1) the risk of complications of migration, protrusion, expulsion and nerve damage resulting from improper insertion and removal of Probuphine and (2) the risks of accidental overdose, misuse and abuse if an implant comes out or protrudes from the skin.

4 PROPOSED REMS FOR PROBUPHINE

The Sponsor's proposed REMS includes a Medication Guide (MG) and elements to assure safe use (ETASU), which include prescriber certification and certification of HCP who dispense (i.e. HCP who Insert/Remove Probuphine). The proposal is described in greater detail below:

4.1 GOALS

The goal of the Probuphine REMS is to mitigate the risk of complications of migration, protrusion, expulsion and nerve damage associated with the improper insertion and removal of Probuphine and the risks of accidental overdose, misuse and abuse if an implant comes out or protrudes from the skin by:

a) Ensuring that prescribers are educated on the following:

- risk of complications of migration, protrusion, expulsion and nerve damage associated with the improper insertion and removal of Probuphine

[5] FDA Efficacy and Safety of Probuphine Clinical Background Memorandum, section 5.1.5.1.1.

[6] FDA Efficacy and Safety of Probuphine Clinical Background Memorandum, section 5.1.5.1.1.

- risks of accidental overdose, misuse and abuse if an implant comes out or protrudes from the skin

b) Ensuring that Probuphine is administered only to patients informed about the risks of complications of migration, protrusion, expulsion and nerve damage associated with the improper insertion and removal, as well as, the risks of accidental overdose, misuse and abuse if an implant comes out or protrudes from the skin.

4.2 MEDICATION GUIDE (MG)

A MG will be provided to each patient prior to the insertion procedure to ensure the patient has been provided adequate information about the potential complications that can arise from the procedure and appropriate wound care. The MG can be used by healthcare providers who insert or remove Probuphine to counsel their patients prior to the procedure.

4.3 ELEMENTS TO ASSURE SAFE USE (ETASU)

4.3.1 Healthcare Provider Certification for Prescriber

Healthcare providers (HCP) who prescribe Probuphine will need to be specially certified in the Probuphine REMS. Certification will include completion of Didactic and Live Practicum Training, as well as, passing the Probuphine REMS Program Knowledge Assessment Test. As a condition of certification, prescribers must counsel patients using the Patient Counseling Tool, ensure that the procedure is only performed under their supervision by a HCP who is certified to insert/remove Probuphine and maintain a copy of the completed Probuphine Insertion/Removal Log in the patient's medical record.

4.3.2 Certification of HCP who Dispense (i.e. HCP who Insert/Remove Probuphine)

To be specially certified to insert/remove Probuphine healthcare providers must complete both the Didactic and Live Practicum Training, as well as, pass the Probuphine REMS Program Knowledge Assessment Test. Additionally, these providers must pass the Live Practicum Training Assessment. The Sponsor has also proposed that in order to become a HCP who inserts or removes the HCP must attest to either having completed appropriate training in a procedural specialty or to having performed a sterile procedure in the 3 months prior to training in the Probuphine REMS.

As a condition of certification, HCPs who insert/remove Probuphine must ensure that the facility where the procedure is being conducted has the appropriate equipment to safely insert/remove Probuphine and that the procedure will take place only in the presence of a certified Probuphine prescriber. Counseling patients using the MG, and also documenting the procedure on the Insertion/Removal Log is also required of these practitioners.

Reviewer Comment:

Under this proposal, only the certified prescriber is able to order and stock Probuphine. A certified HCP who inserts/removes cannot order or stock Probuphine unless they are also a certified prescriber. However, healthcare providers have the option to become

dually certified, that is they are able to certify as both a prescriber and a HCP who inserts/removes. It is important to keep in mind that while the REMS does not specifically require a DATA-2000 waiver to become certified, Probuphine will only be shipped to certified prescribers after verifying that the healthcare provider is DATA-2000 waived.

5 DISCUSSION

5.1 AGENCY COMMENTS ON SPONSOR'S PROPOSAL

The Agency agrees the Sponsor's proposal includes the minimal elements necessary to ensure the benefits outweigh the risks.

The Applicant recognizes that practitioners who have completed a medical residency or fellowship in a procedural specialty may be the most qualified to perform insertions/removals. It is likely that prescribers, who are expected to be specialists in addiction medicine, are likely to have limited experience in performing surgical procedures. Therefore, the REMS will require that all prescribers either (1) become dually certified as HCP who prescribes and inserts/removes Probuphine by taking the live practicum training program and passing the associated assessment of procedural competency (which may be challenging for these practitioners with limited experience performing minor surgical procedures) or (2) arrange for a certified HCP who inserts/removes Probuphine to perform the procedure under the prescriber's supervision.

For HCPs who insert/remove Probuphine, the REMS must include a mandatory, live training program with an assessment of procedural competency. The Sponsor has proposed that HCPs who insert or remove Probuphine must attest to either having completed appropriate training in a procedural specialty or to having performed a sterile procedure in the 3 months prior to training in the Probuphine REMS. However, the REMS with prescriber and dispenser certification provisions generally specify the essential criteria that are necessary to assure safe use of the drug without specifying medical specialties or credentials of HCPs who can or will be permitted to certify. Therefore, the Division of Risk Management believes that HCPs that are able to meet the certification requirements, including passing the assessment of procedural competency, should be able to safely perform the Probuphine procedure. For prescribers who do not insert/remove Probuphine it will require a mandatory training program without an assessment of procedural competency.

Because the REMS will permit a physician to be dually certified as a prescriber and an inserter/remover of Probuphine, it will be possible for a physician who has expertise in performing surgical procedures but does not necessarily regularly manage patients for opioid addiction to certify in the REMS in order to both prescribe (order, stock) and dispense (insert/remove) Probuphine. In this situation, in order to receive Probuphine from a wholesaler, the practitioner would need to be DATA-waived thereby ensuring these practitioners, who may not manage patients for opioid addiction regularly, have completed a DATA-2000 compliant educational program on how to manage opioid addiction.

While DATA-2000 requires the physician who prescribes or dispenses/administers buprenorphine for the treatment of opioid addiction in the office-based setting to be DATA-2000 waived, it permits the DATA-2000 waived physician to allow the

administration of buprenorphine by another individual only if the product is administered in the DATA-2000 waived physician's presence. The proposed REMS permits the administration of Probuphine by a REMS certified HCP who will insert/remove Probuphine in the presence of a REMS certified prescriber who has ordered Probuphine and keeps it in their office stock. The certified prescriber will be responsible for ensuring that the HCP who inserts/removes Probuphine in their presence is certified in the Probuphine REMS Program by contacting the Probuphine REMS Program via the website or call center. Alternatively, a HCP can be dually certified as a prescriber and an inserter/remover of Probuphine. Additionally, although the REMS will not require prescribers to be DATA-waived in order to certify, wholesaler/distributor must verify a physician is DATA-waived in order to ship Probuphine. Therefore, the proposed REMS model complies with the requirements set forth in DATA-2000.

Although not described above, the REMS will include an implementation system, which requires wholesaler/distributors to distribute Probuphine only to certified prescribers, and a timetable for submission of assessments.

5.2 RATIONALE FOR THE REMS REQUIREMENTS

A REMS for Probuphine is required to ensure the benefits outweigh the risks of complications due to improper insertion/removal technique and potential for accidental overdose, misuse, and abuse. The Agency has determined that the REMS include prescriber certification and certification of the HCP who inserts/removes (i.e. administers or dispenses) the medication.

Prescribers may include psychiatrists, primary care physicians and other practitioners in private practice who provide addiction treatment. These prescribers may wish to also perform the insertion and removal procedure for their patients. Because many of these prescribers may have limited experience in performing surgical procedures, the REMS requires mandatory training with an assessment of procedural competency to ensure they have the skills necessary to correctly insert or remove Probuphine. Given the complexity involved in inserting and removing Probuphine, it is yet to be seen how many of these likely prescribers choose to become dually certified as inserters/removers.

Additionally, the REMS includes a patient education component to ensure patients are aware of the serious risk which can occur if the Probuphine implant is expelled. They must be informed of actions to take to prevent further complications and accidental exposures. Both the prescriber and the HCP who inserts the product will be required to counsel the patient on the serious risks associated with Probuphine.

Finally, the REMS includes a closed distribution system that ensures that Probuphine is only distributed to certified prescribers who will ultimately insert the implant or will oversee the insertion of the implant in the patient.

5.3 IMPACT ON BURDEN AND PATIENT ACCESS

The proposed REMS for Probuphine is likely to be burdensome to HCP because it requires certification and training requirements to both prescribe and insert and remove the product. These requirements will likely limit the number of healthcare providers willing to obtain certification.

Additionally, the insertion and removal procedure must take place in an office that is equipped for minor surgical procedures. Not all DATA-waived prescribers who will want to prescribe Probuphine practice in such a location. Even for those prescribers who do work in locations equipped for minor surgical procedures, there will still need to be a certified HCP who inserts and removes on staff or one willing to travel there to perform the procedure. The potential need for 2 separate practitioners to be certified in order to utilize Probuphine for patients is burdensome for the healthcare system.

The proposed Probuphine REMS program will add burden to HCPs which in turn may impact patient access due to HCPs being unwilling to obtain certification. However, the Agency has determined that the proposed REMS is necessary to ensure that the benefits outweigh the risks.

6 CONCLUSION

The proposed REMS is necessary to ensure that the benefits of Probuphine outweigh the potential risk of risks of migration, protrusion, expulsion and nerve damage associated with the improper insertion and removal of Probuphine, as well as, the risks of accidental overdose, misuse and abuse if an implant comes out or protrudes from the skin.

FDA has the authority to require a REMS if additional measures beyond the labeling are necessary to ensure the benefits of a drug outweigh the risks. In considering a REMS for Probuphine, FDA took into account existing requirements for prescribers of buprenorphine under DATA-2000 and applicable provisions under the Controlled Substance Act[7]. The resulting proposed REMS will allow Probuphine to be safely utilized but it is unclear to what extent the burden will have on access to patients who are appropriate candidates for therapy.

[7] Title 21 United States Code Controlled Substance Act Section 802. Definitions. Available at: http://www.deadiversion.usdoj.gov/21cfr/21usc/802 htm Accessed December 14, 2015.

Guidance for Industry

Non-Inferiority Clinical Trials

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 90 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit comments to the Division of Dockets Management (HFA-305), Food and Drug Administration, 5630 Fishers Lane, rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the *Federal Register*.

For questions regarding this draft document contact Robert Temple at 301-796-2270 or Robert O'Neill at 301-796-1700 (CDER), or the Office of Communication, Outreach, and Development (CBER) at 301-800-835-4709 or 301-827-1800.

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)

March 2010
Clinical/Medical

Guidance for Industry

Non-Inferiority Clinical Trials

Additional copies are available from:
Office of Communication
Division of Drug Information, WO51, Room 2201
Center for Drug Evaluation and Research
Food and Drug Administration
10903 New Hampshire Ave.
Silver Spring, MD 20993
Phone: 301-796-3400; Fax: 301-847-8714
druginfo@fda.hhs.gov
http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm
or
Office of Communication, Outreach, and Development
Center for Biologics Evaluation and Research
Food and Drug Administration
1401 Rockville Pike, Rockville, MD 20852-1448
Phone: 800-835-4709 or 301-827-1800
ocd@fda.hhs.gov
http://www.fda.gov/BiologicsBloodVaccines/GuidanceComplianceRegulatoryInformation/default.htm

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
March 2010
Clinical/Medical

TABLE OF CONTENTS

Guidance for Industry[1]
Non-Inferiority Clinical Trials

This draft guidance, when finalized, will represent the Food and Drug Administration's (FDA's) current thinking on this topic. It does not create or confer any rights for or on any person and does not operate to bind FDA or the public. You can use an alternative approach if the approach satisfies the requirements of the applicable statutes and regulations. If you want to discuss an alternative approach, contact the FDA staff responsible for implementing this guidance. If you cannot identify the appropriate FDA staff, call the appropriate number listed on the title page of this guidance.

I. INTRODUCTION

This guidance provides sponsors and review staff in the Center for Drug Evaluation and Research (CDER) and Center for Biologic Evaluation and Research (CBER) at the Food and Drug Administration (FDA) with our interpretation of the underlying principles involved in the use of non-inferiority (NI) study designs to provide evidence of the effectiveness of a drug or biologic.[2] The guidance gives advice on when NI studies can be interpretable, on how to choose the NI margin, and how to analyze the results.

II. BACKGROUND

This guidance consists of four parts. The first part is a general discussion of regulatory, study design, scientific, and statistical issues associated with the use of non-inferiority studies when these are used to establish the effectiveness of a new drug. The second part focuses on some of these issues in more detail, notably the quantitative analytical and statistical approaches used to determine the non-inferiority margin for use in NI studies, as well as the advantages and disadvantages of available methods. The third part addresses commonly asked questions about NI studies and provides practical advice about various approaches. The fourth part includes five examples of successful and unsuccessful efforts to define non-inferiority margins and conduct NI studies.[3]

FDA's guidance documents, including this guidance, do not establish legally enforceable responsibilities. Instead, guidance describes the Agency's current thinking on a subject and should be viewed as recommendations unless specific regulatory or statutory requirements

[1] This guidance has been prepared by the Office of Biostatistics and the Office of New Drugs in the Center for Drug Evaluation and Research (CDER) and the Center for Biologics Evaluation and Research (CBER) at the Food and Drug Administration.

[2] For the purposes of this guidance, all references to *drugs* include both human drugs and therapeutic biologic products unless otherwise specified.

[3] References: in this guidance, reference to methods or studies are not included in the text; rather they are included in a General Reference section and a separate reference section for the examples in the Appendix.

are cited. The use of the word *should* in Agency guidances means that something is suggested or recommended, not that it is required.

III. GENERAL CONSIDERATION OF NON-INFERIORITY STUDIES: REGULATORY, STUDY DESIGN, SCIENTIFIC, AND STATISTICAL ISSUES

A. Basic Principles of a Non-Inferiority Study

1. Superiority Trials versus Non-Inferiority Trials to Demonstrate Effectiveness

FDA's regulations on adequate and well-controlled studies (21 CFR 314.126) describe four kinds of concurrently controlled trials that provide evidence of effectiveness. Three of them — placebo, no treatment, and dose-response controlled trials — are superiority trials that seek to show that a test drug is superior to the control (placebo, no treatment, or a lower dose of the test drug). The fourth kind of concurrent control, comparison with an active treatment (active control), can also be a superiority trial, if the intent is to show that the new drug is more effective than the control. More commonly, however, the goal of such studies is to show that the difference between the new and active control treatment is small, small enough to allow the known effectiveness of the active control to support the conclusion that the new test drug is also effective. How to design and interpret such studies so that they can support such a conclusion is a formidable challenge.

These active control trials, which are not intended to show superiority of the test drug, but to show that the new treatment is not inferior to an unacceptable extent, were once called equivalence trials, but this is a misnomer, as true equivalence (i.e., assurance that the test drug is not **any** less effective than the control), could only be shown by demonstrating superiority. Because the intent of the trial is one-sided (i.e., to show that the new drug is not materially worse than the control), they are now called non-inferiority (NI) trials. But that too, is a misnomer, as guaranteeing that the test drug is not any (even a little) less effective than the control can only be demonstrated by showing that the test drug is superior. What non-inferiority trials seek to show is that any difference between the two treatments is small enough to allow a conclusion that the new drug has at least some effect or, in many cases, an effect that is not too much smaller than the active control.

The critical difference between superiority and NI trials is that a properly designed and conducted superiority trial, if successful in showing a difference, is entirely interpretable without further assumptions (other than lack of bias or poor study conduct); that is, the result speaks for itself and requires no further extra-study information. In contrast, the NI study is dependent on knowing something that is not measured in the study, namely, that the active control had its expected effect in the NI study. This is critical to knowing that the trial had *assay sensitivity* (i.e., could have distinguished an effective from an ineffective drug). A successful superiority trial has, by definition, assay sensitivity. A "successful" NI trial, one that shows what appears to be an acceptably small difference between treatments, may or

may not have had assay sensitivity and may or may not have supported a conclusion that the test drug was effective. Thus, if the active control had no effect at all in the NI trial (i.e., did not have any of its expected effect), then finding even a very small difference between control and test drug is meaningless, providing no evidence that the test drug is effective. Knowing whether the trial had assay sensitivity relies heavily on external (not within-study) information, giving NI studies some of the characteristics of a historical control trial.

FDA regulations have recognized since 1985 the critical need to know, for an NI trial to be interpretable, that the active control had its expected effect in the trial. Thus, 21 CFR 314.126(a)(2)(iv), unchanged since 1985, says:

> If the intent of the trial is to show similarity of the test and control drugs, the report of the study should assess the ability of the study to have detected a difference between treatments. Similarity of test drug and active control can mean either that both drugs were effective or that neither was effective. The analysis of the study should explain why the drugs should be considered effective in the study, for example, by reference to results in previous placebo-controlled studies of the active control drug.

2. Logic of the NI Trial

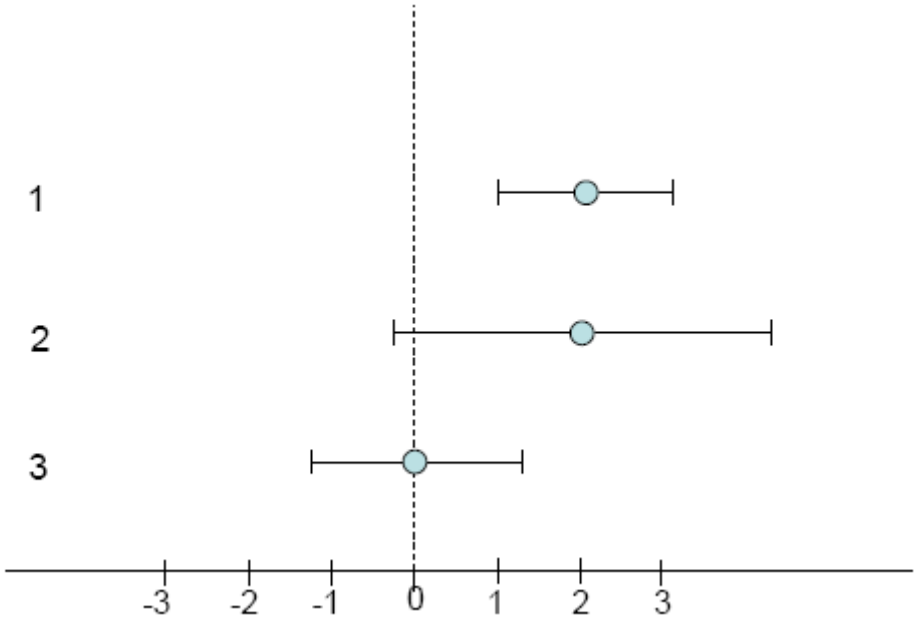
In a placebo-controlled trial, the null hypothesis (H_o) is that the response to the test drug (T) is less than or equal to the response to the placebo (P); the alternative hypothesis (H_a) is that the response to the test drug is greater than P.

$$H_o: T \leq P; \quad T - P \leq 0$$
$$H_a: T > P; \quad T - P > 0$$

In most cases, a treatment effect is established statistically by showing that the lower bound of the two-sided 95% confidence interval (equivalent to the lower bound of a one-sided 97.5% confidence interval) for T-P is > 0.[4] This shows that the effect of the test drug is greater than 0. See Figure 1.

[4] Ref. 4

Figure 1: Three Possible Results of a Placebo-Controlled Superiority Study (Point Estimate, 95% CI)



Test – Placebo (T-P)

1. Point estimate of effect is 2; 95% CI lower bound is 1. Conclusion: Drug is effective and appears to have an effect of at least 1.
2. Point estimate of effect is 2; 95% CI lower bound is <0 (study perhaps too small). Conclusion: Drug is not shown to be effective.
3. Point estimate of effect is 0; 95% CI lower bound is well below 0. Conclusion: Drug shows no suggestion of effectiveness.

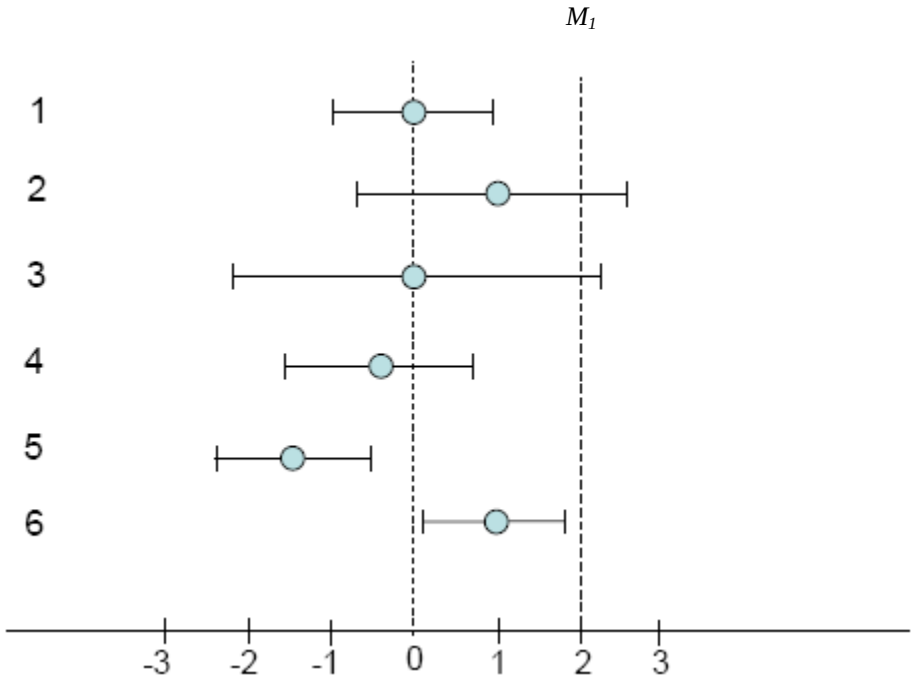
In an NI study whose goal is to show that the new drug has an effect greater than zero, the null hypothesis is that the degree of inferiority of the new drug (T) to the control (C), C-T, is greater than the non-inferiority margin M_1, where M_1 represents what is thought to be the whole effect of the active control (C) relative to placebo in the NI study.[5]

H_o: $C - T \geq M_1$ (T is inferior to the control by M_1 or more)
H_a: $C - T < M_1$ (T is inferior to the control by less than M_1)

[5] M is the non-inferiority margin used in the NI study. It can be no larger than the entire effect that C is presumed to have had in the study, in which case it is called M_1 As described below, the margin of interest can be smaller than M_1, in which case it is called M_2.

Again, non-inferiority is established by showing that the upper bound of the two-sided confidence interval for C-T is < M_1. If the chosen M_1 does in fact represent the entire effect of the active control drug in the NI study, a finding of non-inferiority means that the test drug has an effect greater than 0 (see Figure 2). Thus, in the non-inferiority setting, assay sensitivity means that the control drug had at least the effect it was expected to have (i.e., M_1).

Figure 2: Results of NI Study Showing C-T and 95% CI (M_1 = 2)



1. Point estimate of C-T is 0, suggesting equal effect; upper bound of the 95% CI for C-T is 1, well below M_1; NI is demonstrated.
2. Point estimate of C-T favors C; upper bound of the 95% CI for C-T is >2, well above M_1; NI is not demonstrated.
3. Point estimate of C-T is zero, suggesting equal effect; but upper bound of the 95% CI for C-T is >2 (i.e., above M_1), so that NI is not demonstrated.
4. Point estimate favors T; NI is demonstrated, but superiority is not demonstrated.
5. Point estimate favors T; superiority and NI are demonstrated.
6. Point estimate of C-T favors C and C is statistically significantly superior to T. Nonetheless, upper bound of the 95% CI for C-T<2 (M_1), so that NI is also demonstrated for the NI margin M_1. (This outcome would be unusual and could present interpretive problems.)

The critical problem, and the major focus of this guidance, is determining M_1, which is not measured in the NI study (there is no concurrent placebo group). It must be estimated (really assumed) based on the past performance of the active control and by comparison of prior test conditions to the current test environment (see section III.A.4). Determining the NI margin is the single greatest challenge in the design, conduct, and interpretation of NI trials.

The choice of the margin M_1 has important practical consequences. The smaller the margin, the smaller the upper bound of the 95% two-sided confidence interval for C-T must be, and the larger the sample size that will be needed.

3. Reasons for Using a Non-Inferiority Design

The usual reason for using a non-inferiority active control study design instead of a study design having more readily interpretable results (i.e., a superiority trial) is an ethical one. Specifically, this design is chosen when it would not be ethical to use a placebo, or a no-treatment control, or a very low dose of an active drug, because there is an effective treatment that provides an important benefit (e.g., life-saving or preventing irreversible injury) available to patients for the condition to be studied in the trial. Whether a placebo control can be used depends on the nature of the benefits provided by available therapy. The International Conference on Harmonization guidance E10 on *Choice of Control Group and Related Issues in Clinical Trials* (ICH E10) states:

> In cases where an available treatment is known to prevent serious harm, such as death or irreversible morbidity in the study population, it is generally inappropriate to use a placebo control. [The term "generally" leaves room for a placebo control if the known effective treatment is very toxic.]
>
> In other situations, where there is no serious harm, it is generally considered ethical to ask patients to participate in a placebo-controlled trial, even if they may experience discomfort as a result, provided the setting is non-coercive and patients are fully informed about available therapies and the consequences of delaying treatment.

There are, however, other reasons for using an active control: (1) interest in comparative effectiveness and (2) assessing the adequacy (assay sensitivity) of a placebo-controlled study. These are not the focus of this guidance, but will be considered briefly.

a. Comparative effectiveness

There is growing interest among third party payers and some regulatory authorities, on both cost effectiveness and medical grounds, in the comparative effectiveness of treatments, and an increasing number of such studies are being conducted. A critical issue is the importance of including a placebo group, as well as the active comparator, in such studies (a 3-arm trial) to assess assay sensitivity (i.e., the ability of the trial to detect differences of a specified size between treatments). When the treatment is clinically critical, it will, of course, not be ethically acceptable to include a placebo group, and the discussion of NI studies that follows will be highly relevant to such trials. Even where it would be ethical to include a placebo

group in addition to the active treatments (e.g., in studies of a symptomatic treatment), one is not necessarily included in these comparative trials. Such omission of a placebo group may render such studies uninformative, however, when they show no difference between treatments, unless assay sensitivity can be supported in some other way.

Where comparative effectiveness is the principal interest, it is usually important—where it is ethical, as would be the case in most symptomatic conditions—to include a placebo control as well as the active control. Trials of most symptomatic treatments have a significant failure rate (i.e., they often cannot show the drug is superior to placebo). Where that is the case in a comparative trial, seeing no difference between treatments is uninformative. Inclusion of a placebo group can provide clear evidence that the study did have assay sensitivity (the ability to distinguish effective from ineffective treatments), critical if a finding of no difference between treatments is to be interpretable. For example, we have seen that approximately 50% of all placebo-controlled antidepressant trials of effective agents cannot distinguish drug from placebo. A trial in which two antidepressants are compared and found to have a similar effect is informative only if we know that the two drugs can be distinguished from the concurrent placebo group.

b. Assessing assay sensitivity of a placebo-controlled study

Although a successful superiority trial (e.g., placebo-controlled) is readily interpreted, a failed trial of this design is not. Failure to show superiority to placebo can mean that the drug is ineffective or that the trial lacked assay sensitivity. To distinguish between these two possibilities, it is often useful to include an active control in placebo-controlled studies of drugs in a class or condition where known effective drugs often cannot be distinguished from placebo (e.g., depression, allergic rhinitis, angina, and many other symptomatic conditions). If the active control is superior to placebo but the test drug is not, one can conclude that the test drug lacks effectiveness (or at least is less effective than the active control). If neither the active control nor the test drug is superior to placebo, the trial lacked assay sensitivity and is uninformative about the effect of the test drug.

4. The Non-Inferiority Margin

As described above, the NI study seeks to show that the difference in response between the active control (C) and the test drug (T), (C-T), the amount by which the control is superior to test drug, is less than some pre-specified non-inferiority margin (M). M can be no larger than the presumed entire effect of the active control in the NI study, and the margin based on that whole active control effect is generally referred to as M_1. It is critical to reiterate that M_1 is not measured in the NI trial, but must be assumed based on past performance of the active control, the comparison of the current NI study with prior studies, and assessment of the quality of the NI study (see below). The validity of any conclusion from the NI study depends on the choice of M_1. If, for example, the NI margin is chosen as 10 (because we are sure the control had an effect of at least that size), and the study does indeed rule out a difference of 10 (seeming to demonstrate "effectiveness" of T), but the true effect of C in this study was actually less than 10, say 5, T would not in fact have been shown to have any

effect at all; it will only appear to have had such an effect. The choice of M_1, and assurance that this effect was present in the trial (i.e., the presence of assay sensitivity) is thus critical to obtaining a meaningful, correct answer in an NI study.

Because the consequence of choosing a margin greater than the actual treatment effect of the active control in the study is the false conclusion that a new drug is effective (a very bad public health outcome), there is a powerful tendency to be conservative in the choice of margin and in the statistical analysis that seeks to rule out a degree of inferiority of the test drug to the active control of more than that margin. This is generally done by ensuring that the upper bound of the 95% two-sided confidence interval for C-T is smaller than M_1. The upper bound of the confidence interval for C-T is not, however, the only measurement of interest, just as the lower bound of a 95% confidence interval for effect size of drug versus placebo is not the only value of relevance in a placebo-controlled trial. The point estimate of the treatment effect and the distribution of estimates of C-T smaller than the 95% upper bound are also relevant. Nonetheless, the upper bound of the 95% CI is typically used to judge the effectiveness of the test drug in the NI study, just as a two-sided p-value of 0.05 or less is traditionally the standard used for defining success in a superiority trial. The 95% CI upper bound for C-T is used to provide a reasonably high level of assurance that the test drug does, in fact, have an effect greater than zero (i.e., that it has not lost all of the effect of the active control).

Although the NI margin used in a trial can be no larger than the entire assumed effect of the active control in the NI study (M_1), it is usual and generally desirable to choose a smaller value, called M_2, for the NI margin. Showing non-inferiority to M_1 would provide assurance that the test drug had an effect greater than zero. However, in many cases that would not be sufficient assurance that the test drug had a clinically meaningful effect. After all, the reason for using the NI design is the perceived value of the active control drug. It would not usually be acceptable to lose most of that active control's effect in a new drug. It is therefore usual in NI studies to choose a smaller margin (M_2) that reflects the largest loss of effect that would be clinically acceptable. This can be described as an absolute difference in effect (typical of antibiotic trials) or as a fraction of the risk reduction provided by the control (typical in cardiovascular outcome trials). Note that the clinically acceptable margin could be relaxed if the test drug were shown to have some important advantage (e.g., on safety or on a secondary endpoint).

The definitions used to describe these two versions of M are:

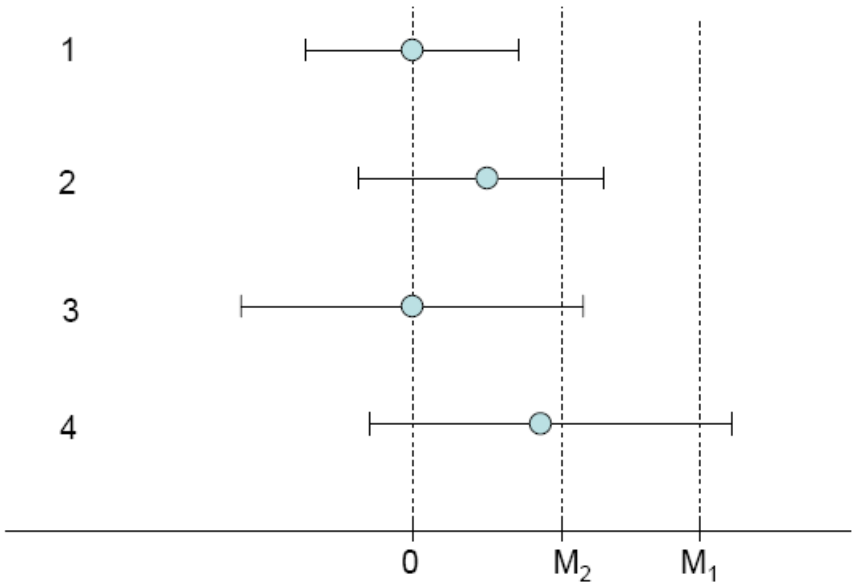
> M_1 = the entire effect of the active control assumed to be present in the NI study
> M_2 = the largest clinically acceptable difference (degree of inferiority) of the test drug compared to the active control

M_1 is based on (1) the treatment effect estimated from the historical experience with the active control drug, (2) assessment of the likelihood that the current effect of the active control is similar to the past effect (the constancy assumption), and (3) assessment of the quality of the NI trial, particularly looking for defects that could reduce a difference between

the active control and the new drug (this diminution of the between-treatment difference is a "bias toward the null" in a trial seeking to show a difference (i.e., superiority), but in this case is a "bias toward the alternative"). Note that because of this third element, the size of M_1 cannot be entirely specified until the NI study is complete.

M_2 is a matter of clinical judgment, but M_2 can never be greater than M_1, even if, for active control drugs with small effects, a clinical judgment might argue that a larger difference is not clinically important. Even if that clinical judgment were reasonable, an M_2 greater than M_1 cannot be used to demonstrate that the test drug has any effect. As explained above, ruling out a difference between the active control and test drug larger than M_1 is the critical finding that supports a conclusion of effectiveness. This analysis is approached with great rigor; that is, a difference (C-T) larger than M_1 needs to be ruled out with a high degree of statistical assurance. As M_2 represents a clinical judgment, there may be a greater flexibility in interpreting a 95% upper bound for C-T that is slightly greater than M_2, as long as the upper bound is still well less than M_1 (see Figure 3).

Figure 3. Active Control – Test Drug differences (Point estimate, 95% CI)



Control – Test (C-T)
(degree of inferiority of test drug)

1. C-T point estimate = 0 and upper bound of 95% CI < M_2, indicating test drug is effective (NI demonstrated).
2. Point estimate of C-T favors C and upper bound of 95% CI < M_1 but > M_2, indicating effect > 0 but unacceptable loss of the control effect.
3. Point estimate of C-T is zero and upper bound of 95% CI < M_1 but it is

slightly greater than M_2. Judgment could lead to conclusion of effectiveness.

4. C-T point estimate favors C and upper bound of 95% CI > M_1, indicating there is no evidence of effectiveness for test drug.

5. Assay Sensitivity and Choosing M_1

Assay sensitivity (AS) is an essential property of a NI clinical trial. AS is the ability of the trial to have detected a difference between treatments of a specified size, M_1 (the entire assumed treatment effect of the active control in the NI trial), if such a difference were present. Stated in another way, AS means that had the study included a placebo, a control drug-placebo difference of at least M_1 would have been demonstrated. As noted, the actual effect of the active control versus placebo is not measured in the NI trial; rather it is estimated (assumed) based on past studies of the drug and comparison of past studies with the current NI study. Note that AS is related to M_1, our best estimate of the effect of the control in the study, even if the NI margin to be used is smaller (M_2). Even if the NI margin to be used is M_2, for example, and is chosen as some percentage of M_1, say 50%, if the active control had an effect of less than M_1 in the trial, the trial would not have shown that M_2 was ruled out.

As noted above, the choice of M_1, and the decision on whether a trial will have AS (i.e., the active control would have had an effect of at least M_1), is based on three considerations: (1) historical evidence of sensitivity to drug effects; (2) the similarity of the new NI trial to the historical trials (the constancy assumption), and (3) the quality of the new trial (ruling out defects that would tend to minimize differences between treatments).

- **Historical evidence of sensitivity to drug effects (HESDE) (ICH E-10)**

HESDE means that appropriately designed and conducted trials in the past that used a specific active treatment (generally the one that is to be used in the new NI study or, in some cases, one or more pharmacologically closely related drugs) regularly showed this treatment to be superior to placebo (or some other treatment). These consistent findings allow for a reliable estimate of the drug's effect size compared to placebo in those past studies, a reasonable starting point for estimating its effect in the NI study. The estimate of effect size must take the variability of past results into account; one would not presume that the largest effect seen in any trial, or even the point estimate of a meta-analysis, is likely to be the effect size in the new study. Analysis of historical data will be discussed further in section IV.

HESDE cannot be determined for many symptomatic treatments, where well-designed and conducted studies often fail to distinguish drug from placebo (e.g., treatments for depression, anxiety, insomnia, angina, symptomatic heart failure, symptoms of irritable bowel disease, and pain). In those cases, there is no reason to assume that an active control would have shown superiority to a placebo (had there been one) in any given NI study, and NI studies of drugs for these treatments are not informative. This is also true for some outcome effectiveness findings, such as secondary prevention of cardiovascular disease with aspirin and post-infarction beta blockade. In the case of aspirin, the largest placebo-controlled trial

(AMIS, the Aspirin Myocardial Infarction Study; see Example 3) showed no effect of aspirin at all, even though other trials all favored aspirin. Similarly, of more than 30 post-infarction beta-blocker trials, only a small number showed significantly improved survival or other cardiovascular benefit.

- **Similarity of the current NI trial to the historical studies – the "constancy assumption"**

The conclusion that HESDE can be used to choose M_1 for the new NI study can be reached only when it is possible to conclude that the NI study is sufficiently similar to the past studies with respect to all important study design and conduct features that might influence the effect size of the active control. This is referred to as the "constancy assumption." The design features of interest include the characteristics of the patient population, important concomitant treatments, definitions and ascertainment of study endpoints, dose of active control, entry criteria, and analytic approaches. The effect of an ACE inhibitor on heart failure mortality has repeatedly been shown in studies where the drugs were added to diuretics and digoxin, but evolution in treatment since those studies were conducted raises questions about our understanding of the present-day effect of these drugs. Since the time of those studies, new medications (beta blockers, spironolactone) have come into standard use. We do not know whether the past effect would still be present when ACE inhibitors are added to a regimen including those two drugs. Similarly, the effect of a thrombolytic on cardiovascular mortality could depend on how soon after symptoms the drug was given, concomitant use of anticoagulants and platelet inhibitors, and use of lipid-lowering drugs. As a general matter, the historical and new NI studies should be as close to identical as possible in all important respects.

It is easier to be reasonably assured that endpoints in the historical trial will be similar to, and will be evaluated similarly to, endpoints in the new trial when these are well-standardized and objective. The effect of the active control could be on a single endpoint (e.g., mortality) or on a composite (e.g., death, heart attack, and stroke), but, again, it is critical that measurement and assessment of these be reasonably consistent over time. The endpoint used in the NI study need not necessarily be the one used in the original trials of the active control if data are available to estimate the occurrence rate of the new endpoint used in the NI study. For example, even if the historical studies used a mortality endpoint, the studies could be used if data could be obtained to calculate an effect size for death plus hospitalization, so long as it was possible to be confident that the circumstances leading to the hospitalization were similar in the historical studies and the NI study. Note, however, that it would not be acceptable to search through a range of endpoints to find the largest historical effect, as this could represent an overestimate of the effect to be expected in the NI study.

In general, where there has been substantial evolution over time in disease definition and treatment, supporting the constancy assumption may be difficult.

Although an NI study can be designed to be similar in most aspects to the historical studies, it may not be possible to assess that similarity fully until the NI study is completed and various

418 characteristics of the study population and response are evaluated. When there is known
419 demonstrated heterogeneity of the active control treatment effect related to patient
420 characteristics (e.g., age, gender, severity), and when that heterogeneity can be quantified, it
421 may be necessary to adjust the estimate of the active control effect size in the NI study if the
422 mix of patient characteristics in the historical and NI studies differ substantially.
423
424 The property of constancy of the treatment effect may depend on which metric is chosen to
425 represent the treatment effect. This issue is discussed in more depth in section IV.B.2.d.
426 Experience suggests that when background rates of outcomes differ among study
427 populations, metrics like hazard ratios or relative risks are more stable than is a metric like
428 absolute effect size, which is more sensitive to changes in event rates in the population.
429
430 • **Good Study Quality**
431
432 A variety of study quality deficiencies can introduce what is known as a "bias toward the
433 null," where the observed treatment difference in an NI study is decreased from the true
434 difference between treatments. These deficiencies include imprecise or poorly implemented
435 entry criteria, poor compliance, and use of concomitant treatments whose effects may overlap
436 with the drugs under study, inadequate measurement techniques, or errors in delivering
437 assigned treatments. Many such defects have small (or no) effects on the variability of
438 outcomes (variance) but reduce the observed difference C-T, potentially leading to a false
439 conclusion of non-inferiority. It should also be appreciated that intent-to-treat approaches,
440 which preserve the principle that all patients are analyzed according to the treatment to which
441 they have been randomized even if they do not receive it, although conservative in
442 superiority trials, are not conservative in an NI study, and can contribute to this bias toward
443 the null. It is more important than usual to plan in advance steps to ensure quality during the
444 conduct of an NI study.
445
446 Finally, it should be recognized that although most investigators seek to carry out high
447 quality trials, the incentives in an NI study are perverse, and quite different from those in
448 superiority trials. In a superiority trial, sloppiness can lead to study failure, and major efforts
449 in trial conduct and monitoring are therefore devoted to avoiding it. In general, sloppiness of
450 any sort obscures true treatment differences. In an NI trial, in contrast, where the goal is to
451 show no difference (or no difference greater than M), poor quality can sometimes lead to an
452 apparent finding of non-inferiority that is incorrect. There is therefore a critical need for
453 particular attention to study quality and conduct when planning and executing an NI study.
454
455 *6. Regulatory Conclusions*
456
457 A successful non-inferiority study shows rigorously that the test drug has an effect greater
458 than zero if it excludes an NI margin of M_1, so long as M_1 is well chosen and represents an
459 effect that the control drug actually would have had (versus a placebo, had there been a
460 placebo group). It can also show that the test drug had an effect greater than some fraction of
461 the control drug effect, depending on the M_2 that is used. It should be appreciated that in
462 addition to the rigorous demonstration of effectiveness, the trial provides additional

information, just as a placebo-controlled trial supporting the effectiveness of a drug does. The point estimate of the drug effect and its confidence interval (usually 95% but could be 90% or 99% under some circumstances) provides information about how large the difference in treatment effect between the test and control drug is likely to be.

In most cases a successful NI study supports effectiveness of the test drug, but it only rarely will support a conclusion that the drug is "equivalent" or "similar" to the active control, a concept that has not been well-defined for these situations. Such similarity might be concluded, however, if the point estimate of the test drug favored it over the control and the upper bound of the 95% CI for C-T was close to showing superiority. Where the chosen M_2 is very small compared to the control drug effect (e.g., a 10% margin in an antibiotic trial in urinary tract infections where response rate is 80%), it might be concluded that the effectiveness of the test drug and control are very similar.

B. Practical Considerations in Use of NI Designs

1. Consider Alternative Designs

ICH E10 identifies a wide variety of study designs that may be better than an NI design in situations where there is difficulty or uncertainty in setting the NI margin, or where the NI margin needs to be so small that the NI study sample size becomes impossibly large.

- **Add-on study**

  In many cases, for a pharmacologically novel treatment, the most interesting question is not whether it is effective alone but whether the new drug can add to the effectiveness of treatments that are already available. The most pertinent study would therefore be a comparison of the new agent and placebo, each added to established therapy. Thus, new treatments for heart failure have added new agents (e.g., ACE inhibitors, beta blockers, and spironolactone) to diuretics and digoxin. As each new agent became established, it became part of the background therapy to which any new agent and placebo would be added. This approach is also typical in oncology, in the treatment of seizure disorders, and, in many cases, in the treatment of AIDS.

- **Identifying a population not known to benefit from available therapy in which a placebo-controlled trial is acceptable**

  In many outcome study settings, effectiveness is established for some clinical settings (e.g., severe disease) but not others. Therefore, it may be possible to study less severely ill patients in placebo-controlled trials. The demonstration that simvastatin was effective in hypercholesterolemic post-infarction patients (4S), for example, did not forestall studies of statins in hypercholesterolemic non-infarction patients (WOSCOPS) or in patients with lesser degrees of hypercholesterolemia (TEXCAPS). This is legitimate so long as one does not in fact know the treatment is of value in the new study population. Recently, it has been possible to study angiotensin receptor

blockers (ARBs) in heart failure in a placebo-controlled trial in patients intolerant of ACE inhibitors (known to improve survival). It would not have been possible to deny a more general population of heart failure patients an ACE inhibitor.

- **Early escape, rescue treatment, randomized withdrawal**

  In symptomatic conditions, there may be reluctance to leave people on placebo for prolonged periods when effective therapy exists. It is possible to incorporate early escape/rescue provisions for patients who do not respond by a particular time, or to use a design that terminates patients on first recurrence of a symptom such as unstable angina, grand mal seizure, or paroxysmal supra- ventricular tachycardia. To evaluate the persistence of effects over time, where conducting a long-term placebo-controlled trial would be difficult, a randomized withdrawal study can be used. Such a study randomly assigns patients treated with a drug for a long period to placebo or continued drug treatment. As soon as symptoms return, the patient is considered to have had an endpoint. This design was first suggested to evaluate long-term benefit in angina.

2. Number of Studies Needed

Ordinarily, with exceptions allowed by the FDA Modernization Act of 1997 (the Modernization Act), FDA expects that there will be more than one adequate and well-controlled study supporting effectiveness. The Modernization Act allows one study plus confirmatory evidence to serve as substantial evidence in some cases, and FDA has discussed in guidance (*Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products*) when a single study might be sufficient.

Where there is uncertainty about the historical effect size (and thus M_1) because of variability or reliance on a single historical study, it will usually be necessary to have more than one NI study to support effectiveness.

Where the studies are of relatively modest size (e.g., most antibiotic NI trials), there is no impediment to conducting more than one NI trial. When the trials needed are very large (to have adequate statistical power), however, this may become a significant problem and it is worth considering what might make a single trial persuasive. Generally, two considerations might do so: (1) prior information, (2) a statistically persuasive result.

- **Prior information**

It is common in NI trials for the test drug to be pharmacologically similar to the active control. (If they were not pharmacologically similar, an add-on study would usually have been more persuasive and more practical). In that case, the expectation of similar performance (but still requiring confirmation in a trial) might make it possible to accept a single trial and perhaps could also allow less conservative choices in choosing the non-

inferiority margin. A similar conclusion might be reached when other types of data are available, for example:

- If there were a very persuasive biomarker confirming similar activity of the test drug and active control (e.g., tumor response, ACE inhibition, or extent of beta blockage)
- If the drug has been shown to be effective in closely-related clinical settings (e.g., effective as adjunctive therapy with an NI study of monotherapy)
- If the drug has been shown to be effective in distinct but related populations (e.g., pediatric versus adult)

- **Statistically persuasive result**

A conclusion that an NI trial can be considered statistically persuasive can be reached in several ways, including the internal consistency of the NI finding, and the margin that is ruled out with a two-sided 95% confidence interval. It is important to recognize that there are two margins of interest, M_1 and M_2. In an NI study, the clinically determined margin M_2 is smaller, often considerably smaller, than M_1, which addresses the question of whether the test drug has any effect. For example, M_2 might be chosen to be 40% of M_1. By meeting this M_2 criterion, ruling out a loss of 40% of the effect of the control, a single NI study provides reasonable assurance that the test drug preserves a clinically sufficient fraction (at least 60%) of the effect of the control treatment. At the same time, it provides strong assurance (probably equivalent in strength to $p \leq 0.001$ in a superiority trial) that the test drug has an effect greater than zero. Particularly where there is strong prior information on the effectiveness of the pharmacological class being studied in the NI trial, showing non-inferiority using M_2 thus provides very strong evidence, analogous statistically to the 2 studies (at $p \leq 0.05$) standard for difference–showing trials, that the new drug has an effect. In such cases, a single such trial would usually be a sufficient basis for approval. Where the effect of the drug is particularly critical, of course, it might be considered necessary to demonstrate that loss of M_2 has been ruled out in more than one study.

In some cases, a study planned as an NI study may show superiority to the active control. ICH E-9 and FDA policy has been that such a superiority finding arising in an NI study can be interpreted without adjustment for multiplicity. Showing superiority to an active control is very persuasive with respect to the effectiveness of the test drug, because demonstrating superiority to an active drug is much more difficult than showing superiority to placebo. Similarly, a finding of less than superiority, but with a 95% CI upper bound for C-T considerably smaller than M_2, is also statistically persuasive.

3. Statistical Inferences

The designer of an NI trial might hope that the test drug is actually superior to the control. It is possible to design the NI study to first test the hypothesis of NI with the pre-specified margin, and then if this test is successful, proceed to analyze the study for a superiority conclusion. This sequential strategy is entirely acceptable. No statistical adjustment is required. A possibility that has thus far had relatively little attention is to have different endpoints with different goals (e.g., superiority on the composite endpoint of death, AMI,

and stroke, but NI on death alone). The multiple endpoints would require some alpha adjustment in such a case, but the procedures here are not well defined. Similarly, if a study had several doses, with interest in NI on each of them and, at the same time, interest in a potential superiority finding for one or more doses, the analytical approach is not yet fully established, although it is clear that some correction for multiplicity would be needed.

Seeking an NI conclusion in the event of a failed superiority test would almost never be acceptable. It would be very difficult to make a persuasive case for an NI margin based on data analyzed with study results in hand. If it is clear that an NI conclusion is a possibility, the study should be designed as an NI study.

4. Choice of Active Control

The active control must be a drug whose effect is well-defined. The most obvious choice is the drug used in the historical placebo-controlled trials. Where studies of several pharmacologically similar drugs have been pooled, which is often done to obtain a better estimate of effect and a narrower confidence interval, and thus a larger M_1, the choice may become complicated. In general, if the drugs in a meta-analysis of placebo-controlled trials seem to have similar effects, any of them could be used as an active control. If their observed treatment effects differ, however, even if not significantly, the one with the highest point estimate of effect should ordinarily be used.

5. Choice of NI Method

The various approaches to calculating the NI margin and analyzing an NI study will be discussed in detail in section IV, but the most straightforward and most readily understood approach will be described here. This method is generally referred to as a fixed margin method and the 95%-95% method (or 90%-95% method, depending on the CIs used to calculate the NI margin) method. The first 95% refers to the confidence interval used to choose the effect size from the historical data, and the second 95% refers to the confidence level used to reject the null hypothesis in the NI study. This approach is illustrated by FDA's evaluation of thrombolytics (TPA). To calculate the NI margin, all available placebo-controlled trials of streptokinase, the active comparator or control, were pooled, giving a point estimate for the effect on survival of a 25% reduction in mortality, with a one-sided 95% lower bound of 22%. As 22% represented the risk reduction by streptokinase compared to placebo, this was translated to the risk increase from being on placebo (1 ÷ .78, or 1.28). The NI study would therefore have had to rule out a 28% increase in risk (the risk increase from a placebo) from not being on TPA. There was a clinical decision to ensure that not more than 50% of the effect of streptokinase was lost, giving an NI margin (M_2) of 1.14, the 95% upper bound of the relative risk for TPA versus streptokinase (see section IV.B.2.c for further discussion of this calculation).

This approach is relatively conservative, as it keeps separate the variability of estimates of the treatment effect in the historical studies and the variability observed in the NI study, and uses a fixed value for the estimate of the control effect based on historical data (the 90% or

95% CI lower bound), a relatively conservative estimate of the control drug effect. On the other hand, a conservative estimate of an important endpoint such as mortality is not necessarily unreasonable, particularly given the uncertainties associated with an NI design.

IV. CHOOSING THE NON-INFERIORITY MARGIN AND ANALYZING THE RESULTS OF AN NI TRIAL

A. Introduction

This section will discuss how to determine the magnitude of the largest acceptable non-inferiority margin, M_1, and the clinical margin, M_2, and how to analyze the NI study. M_1 is the effect the active control (also called positive control) is thought to have had in the NI study. As the effect of the active control in the NI study is not measured (there is no placebo group), this effect must be assumed. The assumed value is based on the analysis of the effect of the active control seen in past controlled studies. M_2 reflects the clinical judgment about how much of M_1 should be preserved by ruling out a loss of M_2. Thus, if it were concluded that it would be necessary for a test drug to preserve 75% of a mortality effect, M_2 would be 25% of M_1, the loss of effect that must be ruled out. It must be appreciated that subjectivity and judgment are involved in all aspects of these determinations, a fundamental difference from a superiority study where all the critical information is measured and no assumptions are needed. This guidance will address how these judgments should be made in selecting the margin selection specified in the NI analysis.

As described in section III, the selection of a margin for an NI study is a two-step process. The first step involves making a reasonable assumption about the effect of the active comparator in the NI study. M_1 is chosen to equal that treatment effect. If the advantage of the control over the test drug in the NI study is larger than M_1, then the test drug has not been shown to have any effect. Effectiveness is therefore demonstrated by showing that the advantage of the control over the test drug (C-T) is smaller than M_1. This can be demonstrated by showing that the upper bound of the 95% CI of C-T is below M_1.

This is very similar to testing a superiority finding at $P \leq 0.05$. If we rule out loss of the entire assumed effect of the control, we can conclude that the test drug is superior to placebo. In most situations where active control studies are used, however, assuring some effect greater than zero is not clinically sufficient, and the second step in selecting the NI margin is choosing a specified portion of the control effect (M_1) whose loss by the test product must be ruled out. This new non-inferiority margin is called M_2, and is based upon clinical judgment. The multiple steps and assumptions that are made in determining an NI margin are all potential sources of uncertainty that may be introduced into the results and conclusions of an NI study. This guidance attempts to identify these sources and suggest approaches to accounting for these uncertainties so that we can reduce the possibility of drawing false conclusions from an NI study.

Conceptually, the NI study design provides two comparisons: (1) a direct comparison of the test drug with the active comparator drug, and (2) an indirect comparison of the test drug to

placebo, based on what is known about how the effect of the active comparator compares to placebo. The entire NI trial concept depends on how much is known about the size of the treatment effect the active comparator will have in the NI study compared to no treatment, but this effect size is not measured in the NI study and must be assumed, based on an analysis of past studies of the control. The validity of the NI trial depends wholly on the accuracy of the assumed effect on the control.

The assumed effect size of the active control in the NI study is based on evidence of that effect derived from past trials, usually trials comparing control with placebo, but trials assessing dose-response, active comparison trials, and even historically controlled trials could play a role. Having assessed the effect of the active control in the past and establishing HESDE (Historical Evidence of Sensitivity to Drug Effect – ICH E-10), it is then necessary to decide whether that effect can be presumed to be present in the new study (the constancy assumption) or must be adjusted in some way based on differences between present-day and historical trials that would reduce the active control effect size. This will be discussed further in section IV.B.2.d. It is also critical to ensure study quality in the NI trial, because poor quality can reduce the control drug's effect size and undermine the assumption of the effect size of the control agent, giving the study a "bias toward the null," which in this case represents the desired outcome.

Having established a reasonable assumption for the control agent's effect in the NI study, there are essentially two different approaches to analysis of the NI study, one called the *fixed margin method* (or the two confidence interval method) and the other called the *synthesis method*. Both approaches are discussed in later sections of section IV and use the same data from the historical studies and NI study, but in different ways.

Briefly, in the fixed margin method, the margin M_1 is based upon estimates of the effect of the active comparator in previously conducted studies, making any needed adjustments for changes in trial circumstances. The NI margin is then pre-specified and it is usually chosen as a margin smaller than M_1 (i.e., M_2), because it is usually felt that for an important endpoint a reasonable fraction of the effect of the control should be preserved. The NI study is successful if the results of the NI study rule out inferiority of the test drug to the control by the NI margin or more. It is referred to as a fixed margin analysis because the past studies comparing the drug with placebo are used to derive a single fixed value for M_1, even though this value is based on results of placebo-controlled trials (one or multiple trials versus placebo) that have a point estimate and confidence interval for the comparison with placebo. The value typically chosen is the lower bound of the 95% CI (although this is potentially flexible) of a placebo-controlled trial or meta-analysis of trials. This value becomes the margin M_1, after any adjustments needed for concerns about constancy. The fixed margin M_1, or M_2 if that is chosen as the NI margin, is then used as the value to be excluded for C-T in the NI study by ensuring that the upper bound of the 95% CI for C-T is $< M_1$ (or M_2). This 95% lower bound is, in one sense, a conservative estimate of the effect size shown in the historical experience. It is recognized, however, that although we use it as a "fixed" value, it is in fact a random variable, which cannot invariably be assumed to represent the active control effect in the NI study.

The synthesis method, derived from the same data, combines (or synthesizes) the estimate of treatment effect relative to the control from the NI trial with the estimate of the control effect from a meta-analysis of historical trials. This method treats both sources of data as if they came from the same randomized trial, to project what the placebo effect would have been had the placebo been present in the NI trial. The process makes use of the variability from both the NI trial and the historical trials and yields one confidence interval for testing the NI hypothesis that the treatment rules out loss of a pre-specified fixed fraction of the control effect, without actually specifying that control effect or a specific fixed NI margin based on the control effect.

B. Statistical Uncertainties in the NI Study and Quantification of Treatment Effect of Active Control

1. What are the Sources of Uncertainty in an NI Study?

There are three major sources of uncertainty about the conclusions from an NI study. Two of these relate to estimating the size of the effect the active control will have in the NI study because that value is the basis for choosing M_1, the non-inferiority margin whose exclusion will be used to conclude that the test drug has an effect. The third is the degree of statistical assurance needed in the NI study itself to determine whether the chosen NI margin has in fact been ruled out.

The first source of statistical uncertainty involves the precision (or variability) of the estimate of the active comparator treatment effect that is derived from an analysis of past data (HESDE), whether this is based on a single randomized active comparator placebo-controlled trial or from multiple trials. The uncertainty of this treatment effect estimate is quantified statistically by using confidence intervals to describe the range within which the true treatment effect size is likely to fall. As described in section III, assurance that the active control will produce a specific effect (at least M_1) in the NI study is the single most critical determination to be made in planning the NI study. Using the point estimate of the treatment effect would not be an acceptable choice for the true treatment effect in the NI study because, on average, half of all trials, even if the historical estimate is correct, would be expected to have a smaller effect, so that one could not be reasonably sure such an effect of the control was present in the NI study. It has therefore become common practice to examine the confidence interval for the effect in historical experience and choose an effect that is reasonably sure to be present in a new study, such as the lower bound of a 95% confidence interval for the historical experience.

Particular problems arise when there is only a single historical study, as there is no information about study-to-study variability (although of course, the confidence interval is likely to be wider when there is only one study), when there are multiple studies but substantial inconsistency in effect sizes among them, and when data from several pharmacologically related drugs are used to develop the estimate for the effect of the active control. When more than a single active comparator study is available, it is necessary to

examine the results from each of the studies to determine whether the treatment effects are consistent among studies or whether there are some studies where the estimate of the treatment effect is zero. The need for some consistency of the active comparator effect size is important and should be considered when choosing M_1. There are also circumstances that might support a less conservative choice for M_1 than the lower bound of the 95% CI for the historical experience. These include factors that strongly support the expectation of a similar clinical effect with the test drug, such as pharmacologic properties of the test drug that are very similar to those of the active control or an effect of the test drug on a persuasive biomarker.

The second source of uncertainty is not statistically based but rather arises from the concern that the effect size estimated from past studies will be different from (larger than) the effect of the active control in the current NI study. The need to assume that the effect will be unchanged is often referred to as the "constancy assumption." If the assumption is incorrect, and the effect size in the current NI study is smaller than the estimated effect from historical studies, M_1 will have been incorrectly chosen (too large) and an apparently successful study showing NI could have given an erroneous result. Lack of constancy can occur for many reasons, including advances in adjunctive medical care, differences in the patient populations, or changes in the assessment of the endpoints under study. As noted in section III, there is some experience to support the view that in outcome studies, the absolute size of the treatment effect is more likely to be variable and sensitive to the background rates in the control group than is the risk reduction. The risk reduction may thus be a more constant (see section IV.B.2.c. on choice of metrics) measure of control drug effect than the absolute effect. How to adjust the NI margin for concerns about constancy is inevitably a matter of judgment.

The third source of uncertainty involves the risk of making a wrong decision from the test of the non-inferiority hypothesis in the NI study (i.e., concluding that $C\text{-}T < M_1$ when it is not). This uncertainty is referred to as the Type I error, or the false positive conclusion risk, and is similar to the concern in a placebo-controlled trial that one might mistakenly conclude that a drug is more effective than placebo. It is, in other words, present in any hypothesis-testing situation. In the NI case, the statistical test is intended to ensure that the difference between control and test drug (C-T, the degree of superiority of the control over the test drug) is smaller than the NI margin, meaning that some of the effect of the control is preserved (if $C\text{-}T < M_1$) or that a sufficient amount is preserved (if $C\text{-}T < M_2$). Typically, the one-sided Type 1 error is set at 0.025, by asking that the upper bound of the 95% CI for C-T be less than the NI margin; this is roughly similar to the usual statistical test for a placebo-controlled trial. If only one NI study is going to be conducted, the probability of a Type 1 error can be made smaller by requiring that the upper bound of a CI greater than 95% be calculated and be less than the margin. This is similar to what is a commonly done for a single placebo-controlled trial (e.g., testing at an alpha of 0.001 instead of 0.05). As noted earlier, however, there may be prior information that eases this concern, and a single study at the usual Type 1 error boundary (0.025) may be considered sufficient if, for example, the drug and active control are pharmacologically similar.

This guidance will discuss the impact of the first two sources of uncertainty on the quantitative approaches to estimating the control treatment effect under different assumptions for these uncertainties, as well as the choice of margin to use in hypothesis testing.

2. Quantification of the Treatment Effect of the Active Comparator

Past controlled studies of the active control provide the empirical data for estimating the size of the treatment effect of the active comparator drug. The magnitude of that treatment effect, which will be the initial basis for determining the control drug effect that can be assumed to be present in the NI study, is critical to determining whether conducting an NI study is feasible. If the active comparator has a small treatment effect, or an effect only marginally distinguished from placebo, or an inconsistent effect, an active controlled study designed to show non-inferiority is likely to require a very large sample size or not be practical at all.

The magnitude of the treatment effect of the active comparator may be determined in several ways, depending upon the amount of data and the number of separate studies of similar design available to support this determination. The availability of many independent studies is generally more informative for this determination, because the estimate of the active comparator treatment effect size can be more precise and less subject to uncertainty, and because it becomes possible to judge the constancy of the effect for at least the period of the studies.

a. Determining HESDE from a single study

The most common situation in which an NI design is used involves outcome studies where the active control drug has been approved for use to reduce the risk of major events (death, stroke, or heart attack). It is not unusual for such approval to have been based on a single study in a specific setting, although there may be other pertinent data in related conditions or in different populations, or with pharmacologically similar drugs. Generally, basing an NI margin on a single randomized placebo-controlled superiority study would need to take into account the variability of the data in that study. The estimate of the treatment effect is usually represented by some metric such as the difference between the event rate in the active treatment group and the placebo control group, which can be an absolute difference in event rates or a risk ratio. The treatment effect has an uncertainty that is usually measured by the confidence interval, a representation of where the result is likely to be 95% of the time (for a 95% CI) in a future study. As a crude gauge, the lower bound of the 95% CI is approximately the effect size demonstrated at a p-value of 0.025 one-sided. It is common to use this value as the effect size we can be reasonably sure the active control had in the historical study and is very likely to have in a future NI study. It is, on average, a low estimate of the effect of the drug, and is "conservative" in that sense, but it is an effect size that has a high probability of being achieved by the active control in the NI study. In contrast, the point estimate of the effect seen in the historical study represents an effect size that may be closer to the true effect of the active control but is one that may not be obtained in a substantial fraction of any new studies. It is critical to choose the estimate of effect size conservatively (i.e., one that previous studies show is very likely to be attained in the NI

study) because the entire logic of the NI study rests on assurance that the active control in the NI study has an effect size at least equal to M_1, the largest possible NI margin.

Generally, therefore, for the fixed margin approach to setting the NI margin, the lower bound of the confidence interval of the effect size of the active comparator in its historical placebo-controlled experience is used to determine M_1 in order to be reasonably sure that the active control will have at least the effect defined as the M_1 in the NI study. The situation improves if the p-value of the estimated treatment effect is much smaller than 0.05, say in the range of 0.01 or 0.001 or even smaller, because in that case the lower bound of the 95% CI will generally be well above zero (in absolute value) or 1.0 (for hazard ratio and other risk estimates). In this case, we are more certain that the treatment effect is real and that the effect of the control in the NI study will be of reasonable size.

When there is only a single trial, there is no objective assessment of study-to-study variability, and there is inevitably concern about the level of assurance we can have that the control will have an effect of a particular size in the NI study. A potential cautious approach to account for this possible variability is to use the lower bound of a wider CI, such as the 99% CI. This is possible where the effect is very large, but will often yield an M_1 that necessitates a very large NI trial. It may be reassuring in such cases if closely related drugs, or the control drug in closely related diseases, have similar effects. A high level of internal consistency in subpopulations (e.g., if the effect of the control drug is similar in subgroups based on gender or age), could also provide some reassurance as to the reproducibility of the result. Such findings might support use of the 95% CI lower bound even if there is only a single study of the active control drug in the population to be studied in the NI trial.

b. Determining HESDE from multiple trials

Identical clinical trials in identical populations can produce different estimates of treatment effect by chance alone. The extent to which two or more studies produce estimates of treatment effect that are close is a function of the sample size of each study, the similarity of the study populations, the conduct of the studies (e.g., dropout rates), and other factors that are probably not measurable. Therefore, another source of uncertainty to be considered when choosing a margin for the current NI study is the study-to-study variability in the estimate of treatment effect.

When there are multiple studies of the active comparator treatment relative to a placebo or no treatment, the opportunity exists to obtain an overall estimate of the active control treatment effect as well as a measure of the study-to-study variability of that treatment effect. When multiple studies of the active control are available, meta-analytic strategies may be used to obtain a more precise estimate of the active control effects. But study-to-study variability in the active comparator treatment effect is a critical consideration as well, because one of the basic assumptions in NI studies is the consistency of the effect size between the historical studies and the current NI study.

Several special cases illustrate the use of multiple studies and problems that can arise. In some of these, when the study-to-study variability is great, the need to provide assurance that the control will have a definable effect size in the NI study (M_1) makes it necessary to adopt a conservative estimate of the effect size.

1. The ideal case is one where there are many studies, each of sufficient size to demonstrate the effect of the active control, or where there are several large outcome studies, each of which has demonstrated an effect of the control, and where the effect sizes derived from these studies are reasonably consistent, so that a pooled estimate, obtained by a meta-analytic approach, provides a very stable and precise estimate of the control effect size (narrow 95% confidence bounds) and allows a choice of M_1 that is large enough to allow a reasonable choice for an M_2 margin and for the design of an NI study of reasonable size.

2. If there are many small studies, where some of them have not demonstrated an effect of the active control, a pooled estimate of the active control effect size and its confidence interval using a random effects model can still be useful, provided there is no evidence of statistical heterogeneity among the study effect sizes.

3. If there are several large outcome studies, some variation of effect sizes is expected, but it would be inappropriate to have the point estimate for one of these fall below the 95% CI lower bound of the pooled study data, suggesting that an explanation of these differences is needed and, in the absence of such an explanation, that it is not possible to determine an NI margin. In this case, a clear failure of one study to show any effect, again, without good explanation, such as wrong choice of endpoint or study population or inadequate sample size, would also argue against the use of an NI design.

4. There are sometimes several large trials of different drugs in a pharmacologic class. Pooling them may allow calculation of a 95% CI lower bound with a narrower CI that yields a higher estimate of the active control drug effect than would any single study. The presumption that the pharmacologically similar drugs would have similar effects may be reasonable, but care should be exercised in extending this assumption too far.

   If the effect size of these different drugs varies considerably in the trials, it may be reasonable to use the pooled data to estimate effect size, but it appears desirable to use the drug with the largest effect (point estimate) as the active control in the NI study, even if the pooled data (95% CI lower bound) are used to estimate the active control effect size.

When an analysis is based on multiple studies, it is important to consider all studies and all patients. Dropping a study that does not show an effect, unless there is a very good reason, can overestimate the control drug effect and give a falsely high M_1. As noted above, the existence of properly designed and sized studies that show no treatment effect of the active comparator may preclude conducting NI studies with that active comparator unless there are valid reasons to explain these results.

Examples 1, 3, and 4 in the Appendix illustrate in more detail how multiple historical placebo-controlled trials of the active comparator studies are evaluated.

c. Metrics of treatment effect

There are several different metrics that can be used to assess the treatment effect estimated in an NI study. These include the following:

- The absolute difference between test and control groups in the proportions of outcomes, cure rates, success rates, survival rate, mortality rate, or the like. This metric is typically used in antibiotic trials.
- The relative risk, or risk ratio (RR), which is the ratio of the rate of events such as death in the treatment and control groups. The risk reduction is 1-RR. Thus, if a treatment has a relative risk of 0.8 compared to placebo, it gives a risk reduction of 20%.
- The hazard ratio is the ratio of the hazards with the test treatment versus the control, much like relative risk, but it is a metric that represents the time specific rate of an event. It is usually employed for time to event or survival type studies.
- The odds ratio is a ratio of the odds of success or survival (or failure/death) of one treatment relative to the other. Note that when event rates are low, as is the case for many cardiovascular outcome studies, risk ratios and odds ratios are quite similar.
- The log of the relative risk, the odds ratio, or the hazard ratio can be used to make the metrics normally distributed and easier to evaluate in the analysis.

The metric used in calculating HESDE need not be the one used in the original study. If placebo response rates differ markedly among several studies in a meta-analysis, it is generally more sensible to analyze relative risk than absolute risk. It seems far more likely that in the NI study it will be the risk reduction, not the absolute effect, that will be constant.

Another consideration that is important for characterizing the treatment effect for time to event studies (which many mortality studies are) is the proportionality of the hazard ratio over the time domain of study treatment exposure. Since the treatment effect is reduced to a single estimated hazard ratio that expresses the treatment effect over the entire time period of exposure, it is important to be aware of and check that the assumption of a proportional or constant hazard ratio is appropriate for the drug and disease situation. The metric that is chosen will determine how the metric behaves in different scenarios, and may be critical in choosing the duration of the NI study.

Note that we are using the convention that for the ratio of risks (bad outcomes such as failure rates or deaths) in the historical trials, risks are shown as control drug/placebo (i.e., the drug is the numerator), so that the RR (or HR) will be less than 1. In an NI study, the control drug becomes the denominator and the test drug is the numerator, with a risk increase to be ruled out. For example, if the control gives a 25% risk reduction relative to placebo, what must be ruled out to show that the NI margin is excluded is an increased risk of 33%, or an RR of

1.33, calculated by dividing the active drug effect versus placebo into 1 (1÷0.75=1.33). How to calculate M_2 is not entirely straightforward. If we take half of the control effect versus placebo, for an HR of 0.875, then convert that to the risk increase to be ruled out, we get 1÷0.875 or 1.14. If, on the other hand, we take half of the 33% increase calculated earlier, we get 1.165.

Whether to calculate M_2 before or after changing numerator and denominator is not settled. A way to calculate the margin without this asymmetry is to convert the HR to the natural logarithm scale. When the natural logarithm transformation of the risk ratio is used, that is, log (A/B) and log (B/A), the two logs have the same magnitude except that the signs are opposite. In the previous example, for 50% retention of the 25% risk reduction in the NI study, the non-inferiority margin for log (T/C) is the mid-point between log (4/3) and zero. By converting log risk ratio back to risk ratio, the non-inferiority margin for T/C is the square root of 4/3, giving a value of 1.155. The margin calculated that way then falls between the 1.14 and 1.165 calculated previously.

The difference between expressing the treatment effect as the absolute difference between success rates in treatment groups and as the relative risk or risk ratio for success on the test treatment relative to the active comparator is illustrated in the following two examples.

For the first example, consider a disease where the cure rate is at least 40% in patients receiving the selected active control and 30% for those on placebo, a 10% difference in cure rates. If the purpose of an NI study is to demonstrate that the test product is effective (i.e., superior to a placebo), then the difference between the test product and active control in the NI study must be less than 10%. The margin M_1 would then be 10%. If the additional clinical objective is to establish that the test product preserves at least half of the active control's effect, then the cure rate of the test product must be shown to be less than 5% worse than the control, the M_2 margin.

This approach depends on the control drug's having an effect of at least 10% greater than a placebo (had there been one) in the NI study. If the population in the NI study did not have such a benefit (e.g., if the patients all had viral illnesses such that the benefit was less than 10%), then even if the 5% difference were ruled out, that would not demonstrate the desired effectiveness (although it would seem to). Note that in this case, if the true effect of the control in the study were 8%, then ruling out a 5% difference would in fact show some effect of the test drug, just not the desired 50% of control effect.

The second example illustrates a non-inferiority margin selected for the risk ratio (test/control) metric. Let C and P represent the true rates of an undesirable outcome for the control and a placebo, respectively. The control's effect compared to placebo is expressed by the risk ratio, C/P. A risk ratio of 1 represents no effect; a ratio of less than 1 shows an effect, a reduction in rate of undesirable outcomes.

Metrics like the risk ratio may be less affected by variability in the event rates in a placebo group that would occur in a future study. For example, a risk ratio for the event of interest of

3/4 =0.75 can be derived from very different absolute success results from different studies, as shown in the table below. While the risk ratio is similar in all four hypothetical studies, the absolute difference in success rates ranges from 5% to 20%. Suppose that the NI margin were based on historical studies showing control drug effects like those in the fourth study. The NI margin would then be chosen as 20%. Now suppose that under more modern circumstances the NI study had a control rate more like Study 1 and an effect size vs. placebo of far less than 20%. An NI margin (M_1) of 20% would then be far greater than the drug effect in the NI study, and ruling out a difference of 20% would not demonstrate effectiveness at all. Thus, if the NI margin were chosen as ruling out an inferiority of 33% (or a relative risk of 1.33, i.e., 1 ÷ 0.75), if the control rate were 15%, the difference (M_1) between test and control would need to be less than 5% (15% x 1.33 = 20%, or 5% > the 15% rate in the active control group).

Study Number	Risk Ratio (C/P)	Control rate	Placebo rate
Study 1	3/4	15%	20%
Study 2	3/4	30%	40%
Study 3	3/4	45%	60%
Study 4	3/4	60%	80%

In this case, where absolute effect sizes vary but risk reductions are reasonably constant, the risk ratio metric provides a better adjustment to the lower event rate in the NI study.

These examples illustrate the importance of understanding how a particular metric will perform. The choice between a relative metric (e.g., risk ratio) and an absolute metric (e.g., a difference in rates) in characterizing the effects of treatments may also be based upon clinical interpretation, medical context, and previous experience with the behavior of the rates of the outcome.

d. The Concept of "Discounting" the Treatment Effect Size to Account for Various Sources of Uncertainty

One of the strategies employed in choosing the margin M_1 for the NI study design is that of "discounting" or reducing the magnitude of the margin size that is used in the NI study from what is calculated from the analysis of HESDE. Such discounting is done to account for the uncertainties in the assumptions that need to be made in estimating, based on past performance, the effect of the active control in the NI study. This concept of discounting focuses on M_1 determination and is distinct from a clinical judgment that the effect that can be lost on clinical grounds should be some fraction of M_1 (i.e., M_2). As discussed above, there are uncertainties associated with translating the historical effect of the active control (HESDE) to the new situation of the active control NI trial, and it is tempting to deal with that uncertainty in the constancy assumption by discounting the effect ("take half"). To the extent possible, concerns about the active control effect should be as specific as possible, should use available data (e.g., magnitude of possible differences in effect in different patient population, consistency of past studies, and consistency within studies across population subsets should be examined), and should take into account factors that reduce the need for a

conservative estimate, such as the pharmacologic similarity of the test and control drugs and pharmacodynamic effects of the new drug, rather than reflecting "automatic" discounting. Having considered these matters, if significant uncertainties remain, an approach that further discounts or reduces, say by 25%, the magnitude of the active control effect based on HESDE can be considered.

A closely related issue is adjustment of M_1 to reflect a finding that the population in the NI study was different from the historical study in such a way that what the historical experience shows would lead to a smaller effect size (e.g., a finding of a smaller effect in women would need to be considered in assessing the validity of M_1 if the NI study had substantially more women than the historical studies). In general, the assessment of the historical data should identify such differences so that plans for the NI study take this into account or so that the value of M_1 can be revisited in light of the study population included in the NI study.

C. Statistical Methods for NI Analysis

Several approaches are used to demonstrate statistically that the NI objective is met. Each statistical approach to demonstrating NI depends upon a number of factors including:

- What assumptions are made and how verifiable or empirically demonstrable these assumptions are
- The degree to which judgment, both statistical and clinical, is exercised in accounting for the various uncertainties in the data from the current NI study and also from the clinical trials of the active control that are the basis for estimating its effect
- The clinical judgment of how much of the treatment effect of the active comparator can be lost (M_2 selection)

As noted earlier, the two main approaches to demonstrating non-inferiority are the fixed margin method and the synthesis method.

Each of these statistical approaches uses the same data from the previously conducted controlled trials of the active control and the same data from the current NI study, but the approaches are different in several ways. The first is with regard to their emphasis on the specific determination for M_1 before determining M_2. There is also a difference between them in how the data from the historical studies and the NI study are used or combined. What follows is a guide to the differences between the two approaches. Examples 1(A) and 1(B) in the Appendix provide more detailed illustrations of how each of these approaches is used and interpreted. In general, the fixed margin approach is more conservative and treats the variance of the NI study and historical evidence distinctly. That is, a very large historical database will give a narrower CI and larger 95% lower bound for M_1, but it will not directly figure into the test drug versus placebo calculation, as is done in the synthesis method. Concern about using the synthesis approach reflects our view that the method incorporates too much certainty about the past results into the NI comparison. We believe the fixed margin approach is preferable for ensuring that the test drug has an effect greater than

placebo (i.e., the NI margin M_1 is ruled out). However, the synthesis approach, appropriately conducted, can be considered in ruling out the clinical margin M_2.

1. The Fixed Margin Approach for Analysis of the NI Study

Sections IV.B.2.a and B.2.b contain discussions of the basic statistical approach to estimating the active comparator treatment effect size from past controlled trials. The goal of these analyses is to define the margin M_1, a fixed value, based on the past effect of the active control, which is intended to be no larger than the effect the active control is expected to have in the NI study. Whether M_1 is based on a single study or multiple studies, the observed (if there were multiple studies) or anticipated (if there is only one study) statistical variation of the treatment effect size should contribute to the ultimate choice of M_1, as should any concerns about constancy. The selection of M_2 is then based on clinical judgment regarding how much of the M_1 active comparator treatment effect can be lost. The exercise of clinical judgment for the determination of M_2 should be applied after the determination of M_1 has been made based on the historical data and subsequent analysis.

All relevant studies of the active comparator and all randomized patients within these studies should generally be used in determining the margin M_1 because that provides a more reliable and, possibly, conservative estimate. The actual selection of which studies are used in a meta-analysis and how that selection is made can be complex and itself subject to judgment. See Examples 1(A), 3, and 4 that illustrate these points in the Appendix.

The design and analysis of the NI study, and its analysis using the fixed margin approach, is well known and described in ICH E9, section 3.3.2. This statistical approach relies upon the choice of a fixed non-inferiority margin that is pre-specified and part of the NI design. There is very little, however, in ICH E9 or ICH E10 that discusses just how to determine the margin. Although the constancy assumption and study quality issues are recognized, there is little discussion about how to adjust the margin because of such statistical or study data uncertainties. Any discounting of the historical evidence of the effect of the active control based on uncertainty of the constancy of the effect (e.g., because of changes in practice or concomitant treatment), which is an attempt to improve the estimate of the control effect in the NI study, affects the M_2 as well, as in most cases M_2 is a fraction of M_1. M_2 might not be affected when it is very small compared to M_1, as is the case in considering very effective drugs. It is critical to note that M_2 is a judgment that is made after M_1 is chosen, but M_2, of course, can never be larger than M_1. It is perhaps tempting to make up for uncertainty in M_1 by demanding assurance of preservation of a larger fraction of M_1 by ruling out a smaller loss of effect (i.e., using a smaller M_2), but the temptation should be avoided. The first and most critical task in designing an NI study is obtaining the best estimate of the effect of the active control in the NI study (i.e., M_1).

Operationally, the fixed margin approach usually proceeds in the following manner. The active comparator effect size is calculated from past placebo-controlled studies. The lower bound of the confidence interval describing the effect of the active control in past studies, a single number, is selected as a conservative choice for the active comparator effect size.

While traditionally the 95% confidence interval is used, there can be flexibility in this choice, such as a 90% confidence interval or even narrower, when the circumstances are appropriate to do so (e.g., strong evidence of a class effect, strong biomarker data). It is recognized that use of a fixed margin to define the control response is conservative as it picks a "worst case" out of a confidence interval that consists of values of effect that are all larger. This choice, however, is one response to the inherent uncertainty of estimates based on past studies, including the variability of those past estimates, and the possibility that changes in medical practice, or hard to recognize differences between the past studies and the current NI study, have made the past effect an overestimate of the active control effect in the new study.

Although some of the uncertainty about applicability of past results to the present is reflected in a conservative choice of margin (95% of CI lower bound) used to initiate consideration of M_1, there may be further concerns about past variability and constancy that lead to a determination to discount this lower bound further in choosing M_1 to account for any sources of uncertainty and dissimilarities between the historical data and the NI study to be conducted, as discussed in the earlier sections. Following this, a clinical judgment is made as to how much of this effect should be preserved. This clinical judgment could choose M_2 to be the same as M_1, but as noted, where the treatment effect is important (e.g., an effect on mortality) it is usual to ask that a reasonable fraction of the control effect be preserved, by making M_2, the loss of effect to be ruled out, smaller than M_1. Choosing M_2 as 50% of M_1 has become usual practice for cardiovascular (CV) outcome studies, whereas in antibiotic trials, where effect sizes are relatively large, a 10-15% NI margin for M_2 is common. Note that the M_2 of 50% of M_1 is on a relative scale, whereas the 10-15% is on the absolute scale for antibiotic drugs. The analysis of the NI study involves only the data from the NI study, and the test of the hypothesis that inferiority greater than the M_2 margin has been excluded is statistically similar to showing that the 95% CI in a superiority study excludes a difference of zero.

Thus, there are two confidence intervals involved in the fixed margin approach, one from the historical data, where one uses the lower bound to choose M_1, and one from the NI study (to rule out C-T > M_2); in this example both intervals are 95% confidence intervals. That is why this fixed margin approach is sometimes called the 95%-95% method. It should be appreciated that the analysis of the NI study (ruling out a difference > M_2 by examining the lower bound of the CI for C-T) is the analysis that is based on the randomized comparison in the NI study, in contrast to the determination of M_1, which is not based on a concurrent randomization.

Separating the process of estimating the treatment effect of the active comparator based upon the historical data (i.e., choice of M_1) from the analysis of the NI study has some advantages and disadvantages. Two important advantages are that it provides a single number that is clinically understandable for an M_1 (and derived M_2) and that it provides a basis for planning the sample size of the NI study to achieve statistical control of Type 1 error and the power needed for the NI study to meet its objective for the pre-specified NI margin. One arguable disadvantage is that the method is statistically not efficient because it uses the two confidence interval approach rather than a combined estimate of the statistical variability of the historical

and NI study data. Nevertheless, use of the fixed margin is readily understood, particularly by non-statisticians, and is only somewhat conservative compared to an analysis using the synthesis approach. Decisions to discount the M_1 further or, where appropriate, to use a narrower confidence interval, are easily explained, and can make the fixed margin approach more or less conservative.

Deciding on the NI clinical margin M_2 is also a relatively straightforward concept. It is plainly a matter of judgment about how much of the treatment effect must be shown to be preserved, a consideration that may reflect the seriousness of the outcome, the benefit of the active comparator, and the relative safety profiles of the test and comparator. It also has major practical implications. In large cardiovascular studies, it is unusual to seek retention of more than 50% of the control drug effect even if this might be clinically reasonable, because doing so will usually make the study size infeasible.

The fixed margin approach considers the NI margin as a single number, fixed in advance of the NI study. The hypothesis tested in the NI study determines whether the comparison of the test drug to the active control meets the specified NI criterion, assuming, of course, that the active control had at least its expected effect (equal to M_1) and that the study therefore had assay sensitivity. A successful NI conclusion, ruling out a difference > M_1, shows that the test drug is effective (just as a superiority study showing a significant effect at $p \leq 0.05$ does) and, if a difference > M_2 is also ruled out, shows that the new drug preserves the desired fraction of the control drug's effect. This statistical test of hypothesis is not formally directed at determining whether the test drug would have been superior to a placebo, had a placebo group been included in the NI study, but it leads to a similar conclusion by ruling out the possibility that the test drug is inferior to the control by more than an amount equal to the whole effect of the control compared to placebo (that effect being known from past studies).

The possible outcomes of such trials are shown in Figures 2 and 3 in section III of this guidance.

2. The Synthesis Approach for Analysis of NI

An alternative statistical approach is known as the synthesis approach because it combines or synthesizes the data from the historical trials and the current NI trial, reflecting the variability in the two data sets (the current NI study and the past studies used to determine HESDE). The synthesis method is designed to directly address the question of whether the test product would have been superior to a placebo had a placebo been in the NI study, and also to address the related question of what fraction of the active comparator's effect is maintained (the loss to be ruled out) by the test product. In the synthesis approach, the NI margin is not predetermined, but the outcome of the NI study, a consideration of the effect of the test agent vs. placebo, can be judged for adequacy.

Although the synthesis approach combines the data from the historical trials into the comparison of the concurrent active comparator and the test drug in the NI study, a direct randomized concurrent comparison with a placebo is of course not possible, as the placebo

group is not a concurrent control and there is no randomization to such a group within the NI study. The imputed comparison with a placebo group that is not in the NI study thus rests on the validity of several assumptions, just as the fixed margin approach does. The critical assumption of the constancy of the active control effect size derived from the historical controlled trials is just as important when the synthesis method is used.

Because of the way the variance of the historical data and the NI data are combined for the synthesis test, the synthesis test is more efficient (uses a smaller sample size or achieves greater power for the same sample size) than the fixed margin approach but requires assumptions that may not be appropriate. The statistical efficiency of the synthesis approach derives primarily from how the standard error of the comparison of test product to active comparator is dealt with. See Appendix, Example 1(B), for a comparison of the two methods and the variance calculations.

The synthesis approach does not specify a fixed NI margin. Rather, the method combines (or synthesizes) the estimate of treatment effect relative to the control from the NI trial with the estimate of the control effect from a meta-analysis of historical trials. The method treats both sources of data as if they came from the same randomized trial, to project where the placebo effect would have been had the placebo been present in the NI trial. The synthesis process makes use of the variability from the NI trial and the historical trials and yields one confidence interval for testing the NI hypothesis that the treatment preserves a fixed fraction of the control effect, without actually specifying that control effect or a specific fixed NI margin based on the control effect. Clinical judgment is used to pre-specify an acceptable fraction of the control therapy's effect that should be retained by the test drug, regardless of the magnitude of the control effect.

A disadvantage of the synthesis approach, however, is that it does not allow for a pre-specification of the actual size or magnitude of the NI margin M_1, so the clinical judgment to determine the choice of M_2 is difficult and is generally not made until results are seen. Moreover, it may be unrealistic to assign the same weight to the variance of the historical outcome data and to that of the concurrent randomized NI treatment As also noted, the efficiency of the fixed margin approach can sometimes be enhanced either formally, by including more trials (e.g., of related drugs) in the historical meta-analysis, and thereby increasing the margin M_1, or, as a matter of judgment, by considering pharmacologic similarities between the control and test drugs, effects on pertinent biomarkers (e.g., tumor response rate), all of which could lead to choice of a fixed margin based on a less extreme boundary of the confidence interval (e.g., 80% instead of 95%).

D. Considerations for Selecting M_2, the Clinical Margin, and the Role of Subjective Judgment

M_2 is the margin that is the pre-specified NI margin that should be met in an NI study. The determination of M_2 is based on clinical judgment and is usually calculated by taking a percentage or fraction of M_1. The clinical judgment in determining M_2 may take into account the actual disease incidence or prevalence and its impact on the practicality of sample sizes

that would have to be accrued for a study. There can be flexibility in the M_2 margin, for example, when:

(1) The difference between the active comparator response rate and the spontaneous response rate is large;
(2) The primary endpoint does not involve an irreversible outcome such as death (in general, the M_2 margin will be more stringent when treatment failure results in an irreversible outcome);
(3) The test product is associated with fewer serious adverse effects than other therapies already available;
(4) The test product is in a new pharmacologic category and has been shown to be tolerated by patients who do not tolerate therapies that are already available.

There is also a difference in implication when the study NI conclusion is "not quite" significant (M_1 is not excluded) for M_1 and when this is the case for M_2. Failure to exclude inferiority equal to M_1 means there is no assurance of any effect. Just as, for a placebo-controlled trial, it would be most unusual to accept as positive a study with $p > 0.05$, it would be most unusual to accept an NI study where the upper bound of 95% CI was $> M_1$. On the other hand, failing to exclude M_2 by a small amount means that instead of ruling out a loss of 50% of M_1, you have ruled out, say, a 48% loss, not necessarily a definitive failure. As noted above, we would also consider the less conservative synthesis approach in assessing M_2.

E. Estimating the Sample Size for an NI Study

It is important to plan the sample size for an NI clinical trial so that the trial will have the statistical power to conclude that the NI margin is ruled out if the test drug is truly non-inferior. This is not always an easy task. At the protocol planning stage, using the fixed margin approach, the magnitude of the NI margin will be specified; the sample size must be based on the need to rule out inferiority greater than M_2. This should usually be based on an NI using a fixed margin approach. The margin to be ruled out is the most critical component of the sample size planning, but the variance of the estimate of the treatment effects will not be known and it is also critical. A further problem is posed by the possibility that event rates will be lower in the new study. In this case, if the NI margin is expressed as, for example, ruling out (at the upper bound of the 95% CI for C-T) an increase in risk of 25%, this will be far easier when the event rate on active control is 8% than when it is 4%, even if the active control is superior to placebo by the same absolute 20% difference. This problem is not different from specifying sample size in a superiority trial. It too depends on the event rate, and it is common to examine blinded data during the trial to see if the event rate is unexpectedly low. A similar approach could be applied in an NI trial with upward adjustment of the sample size if the event rate is unexpectedly low. There is one further consideration. If, in reality, the test drug is somewhat more effective than the control, it will be easier to rule out any given NI margin and a smaller sample size could be used. A somewhat less effective test drug will, of course, require a larger sample size.

F. Potential Biases in an NI Study

Traditionally, analysis of the results of randomized clinical superiority trials follows the intent-to-treat principle, namely, that all randomized patients are analyzed according to the treatment to which they were randomized. This analysis is intended to avoid various biases associated with patients switching treatment, selection bias, and dropout/withdrawal patterns that may confound the observed treatment effect. This is recognized as a potentially conservative analysis. Including patient outcomes that occur after a patient has stopped the treatment, for example, or show poor compliance with treatment, would be expected to bias the analysis toward the null (no treatment difference). Intent-to-treat)ITT) analyses in superiority trials are nonetheless preferred because they protect against the kinds of bias that might be associated with early departure from the study. In non-inferiority trials, many kinds of problems fatal to a superiority trial, such as non-adherence, misclassification of the primary endpoint, or measurement problems more generally (i.e., "noise"), or many dropouts who must be assessed as part of the treated group, can bias toward no treatment difference (success) and undermine the validity of the trial, creating apparent non-inferiority where it did not really exist. Although an "as-treated" analysis is therefore often suggested as the primary analysis for NI studies, there are also significant concerns with the possibility of informative censoring in an as-treated analysis. It is therefore important to conduct both ITT and as-treated analyses in NI studies. Differences in results using the two analyses will need close examination. The best advice for conducting an NI study is to be aware at the planning stage of these potential issues and to monitor the trial in a manner that minimizes these problems, as they can seriously affect the validity of an NI study.

Other sources of bias that could occur in any study are also of concern in the NI study and are of particular concern in an open label study. For such open label NI studies, how best to ensure unbiased assessment of endpoints, unbiased decisions about inclusion of patients in the analysis, and a wide variety of other potential biases, need particular attention.

G. Role of Adaptive Designs in NI Studies — Sample Size Re-estimation to Increase the Size of an NI Trial

Because it may be difficult to adequately plan the sample size for any study, including an NI study, especially when assumptions like the event rate may change from the planning phase to the study conduct, adaptive study designs that can allow for the prospective re-estimation of a larger sample size can be considered. The most critical single consideration in such designs is precise knowledge about whether there is unblinding as to treatment. Sample size re-estimation, if based on a blinded analysis of the overall variance estimate or the overall event rate, without knowledge of or a comparison of the unblinded treatment group response rates or the differences between treatment groups, is not only acceptable but generally advisable. It is critical to provide reassurance and procedures that ensure maintenance of blinding.

If an adaptive design that allows unblinding is contemplated, then the design features and procedures for protection of the integrity of the trial need to be clearly stated in the protocol

for the trial. Some adaptive designs may include an independent Data Monitoring Committee (DMC) to monitor the planned adaptation. The DMC charter should address procedures for the sharing and blinding of data, and the procedures used to maintain a firewall between those who do, and those who do not view unblinded data. Some of these issues will be addressed in a companion guidance on Adaptive Study Designs.

H. Testing NI and Superiority in an NI Study

In general, when there is only one endpoint and one dose of the test treatment, a planned NI study can be tested for superiority without a need for Type 1 error alpha correction. That is, the same 95% or higher confidence interval employed for testing non-inferiority with the pre-specified fixed margin can be used to test superiority. One can also think of this as a two-stage analysis in which the showing of NI using a 95% confidence interval (invariably successful if the test drug is actually superior), is then followed sequentially by superiority testing. This sequential testing has the Type I error rates for both non-inferiority and superiority controlled at a level of no more than 5%. A non-inferiority showing after a failed superiority study, in contrast, gives a generally uncertain result, and such a study would generally be considered a failed study. Thus, successful showing of non-inferiority allows superiority testing but a failed showing of superiority would yield credible evidence of non-inferiority only if the study were designed as a non-inferiority study (e.g., the NI margin must be pre-specified, and assay sensitivity and HESDE must be established).

When there are multiple endpoints or multiple doses of the test treatment evaluated in an NI study, the valid statistical decision tree can be very complex. Using the same 95% confidence interval to test non-inferiority and superiority at each endpoint level or at each dose may inflate the overall Type I error rate associated with drawing one or more false conclusions from such multiple comparisons, regardless of whether they are non-inferiority or superiority testing. Thus, for any statistical decision tree composed of tests of superiority and non-inferiority in multiple comparison settings, it is imperative to evaluate the overall Type I error rate for all the comparisons involved in the testing and make appropriate statistical adjustments.

Some of the problems in interpreting the results of non-inferiority analyses are more subtle than those with superiority testing. In particular, as noted previously, design or conduct problems such as medication non-compliance or misclassification/measurement error, errors that would be fatal to success in a superiority study, can lead to apparently favorable (results) in a non-inferiority study.

V. COMMONLY ASKED QUESTIONS AND GENERAL GUIDANCE

1. Can a margin be defined when there are no placebo-controlled trials for the active control for the disease being assessed?

If the active control has shown superiority to other active treatments in the past, the difference demonstrated represents a conservative estimate of HESDE, one that can certainly serve as a basis for choosing M_1. It may also be possible that trials of the active control in related diseases are relevant. The more difficult question is whether historical experience from nonconcurrently controlled trials can be used to define the NI margin. The answer is that it can, but the circumstances are similar to those in which a historically controlled trial can be persuasive (see ICH E-10). First, there should be a good estimate of the historical spontaneous cure rate or outcome without treatment. Examination of medical literature and other sources of information may provide data upon which to base these estimates (e.g., historical information on natural history or the results of ineffective therapy). Second, the cure rate of the active control should be estimated from historical experience, preferably from multiple experiences in various settings, and should be substantially different from the untreated rate. For example, if the spontaneous cure rate of a disease is 10-20% and the cure rate with an active control is 70-80%, these are substantially different and an acceptable margin, generally chosen conservatively, can probably be identified for M_1. The clinically acceptable loss of this effect can then be determined for M_2. Estimates of the cure rate of the active control should be based upon data from clinical trials, even if these are not controlled, and it is critical to be sure the trial patients and untreated patients are similarly defined and selected. Example 2 in the Appendix illustrates a case of this kind, in which it was concluded that a margin could be defined despite the absence of placebo-controlled trials of the active control. It becomes more difficult to identify a margin when the difference between the spontaneous cure rate and active drug cure rate is smaller. For example, if the historical spontaneous cure rate is 40% and the active control rate is 55%, it would not be credible to identify the NI margin in this case as 15%, as such a small difference could easily be the result of different disease definition or ancillary therapy. When the historical cure rates for the active control and the cure rate in patients who receive no treatment are not known at all from actual studies (i.e., are just based on clinical impressions), it will be difficult or impossible to define an NI margin.

2. Can the margin M_2 be flexible?

As indicated in sections III and IV, there is a critical difference between demonstrating in the NI study that the margins M_1 and M_2 have been met. M_1 is used to determine whether the NI study shows that the test drug has any effect at all. Accepting a result in which the 95% CI did not rule out loss of M_1 would be similar to accepting, as showing effectiveness, a superiority study whose estimated treatment effect was not significant at $p \leq 0.05$. M_2, in contrast, represents a clinical judgment about what level of loss of the active control effect is acceptable. A typical value for M_2 is often 50% of M_1, at least

partly because the sample sizes needed to rule out a smaller loss become impractically large. In this case, there is a better argument for some degree of flexibility if the study did not quite rule out the M_2 margin; there might be reason to consider, for example, assurance of 48% retention (but not the expected 50%) for M_2 as acceptable. We have also concluded that the fixed margin method, more conservative but with fewer assumptions, should generally be used in ensuring that loss of M_1 is ruled out but that the synthesis method can be used to assess M_2. Of course, allowing too much inferiority of the test drug to the standard, especially for endpoints of mortality and serious morbidity, would clearly not be acceptable.

3. Can prior information or other data (e.g., studies of related drugs, pharmacologic effects) be considered statistically in choosing the NI margins or in deciding whether the NI study has demonstrated its objective?

Prior information could be characterized in a statistical model or in a Bayesian framework by taking into account such factors as evidence of effects in multiple related indications or on many endpoints. Such information might be used in determining M_1 in a more flexible (less conservative) manner. For example, if multiple studies provide very homogeneous results for one or more important endpoints it may be possible to use the 90% lower bound rather than the 95% lower bound of the CI to determine the active control effect size. Similarly, if there were additional supporting evidence for the clinical effect of the test drug, such as prior information on the efficacy of the test drug in related diseases or in a compelling animal model, or an effect on an important biomarker (e.g., tumor response rate), or evidence that pharmacologically related drugs were clearly effective in the condition being studied, such prior information would increase the evidence for the plausibility of the intended NI effect of the test drug, which might allow use of a less conservative estimate of effect than the 95% lower bound of the confidence interval for C-T in the NI study. Finally, a statistical model such as a regression adjustment may be applied to the NI study analysis if the covariates for patients in the historical clinical studies are distributed differently from those of patients in the current NI study. This adjustment may, in some situations, reduce the variance of the NI test and increase the ability of the comparison to meet the NI margin. In other situations, where there is more heterogeneity of the covariates, the variance may be increased, adversely impacting the comparison.

4. Can a drug product be used as the active comparator in a study designed to show non-inferiority if its labeling does not have the indication for the disease being studied, and could published reports in the literature be used to support a treatment effect of the active control?

The active control does not have to be labeled for the indication being studied in the NI study, as long as there are adequate data to support the chosen NI margin. FDA does, in some cases, rely on published literature and has done so in carrying out the meta-analyses of the active control used to define NI margins. An FDA guidance for industry on *Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products*

describes the approach to considering the use of literature in providing evidence of effectiveness, and similar considerations would apply here. Among these considerations are the quality of the publications (the level of detail provided), the difficulty of assessing the endpoints used, changes in practice between the present and the time of the studies, whether FDA has reviewed some or all of the studies, and whether FDA and the sponsor have access to the original data. As noted above, the endpoint for the NI study could be different (e.g., death, heart attack, and stroke) from the primary endpoint (cardiovascular death) in the studies if the alternative endpoint is well assessed (see also question 6).

5. If the active control drug is approved for the indication that is being studied, does the margin need to be justified, or if the active control drug has been used as an active comparator in the past in another study of design similar to the current study and a margin has been justified previously, can one simply refer to the previous margin used?

When an active control drug is approved, the effect size for the indication is not usually identified in a pooled analysis, nor is the variability of that effect size in the various trials calculated. It would therefore be difficult to base the NI margin on the label of the active control drug. On the other hand, FDA's reliance on the studies for approval would support the view that the quality of the studies was acceptable and that the studies could contribute to a determination of the NI margin. In general, approval of a drug is based on showing superiority to placebo, usually in at least two studies, but FDA may not have critically assessed effect size and may not have closely analyzed "failed" studies. In general, FDA will usually not have carried out a meta-analysis of the trials. It is therefore essential to use the data from all available controlled trials (unless a trial has a significant defect), including trials conducted after marketing, to calculate a reasonable estimate of the actual control effect size, as described above. If the active-control data have been used to define a NI margin for another study, it is important to determine that the previous conclusion is applicable to the new study, but in general such prior use should indicate that FDA has assessed the NI margin for a NI study with similar endpoints and population.

6. What are the choices of endpoints to be aware of before designing a non-inferiority trial design?

The endpoints chosen for clinical trials (superiority or NI) reflect the event rate in the population, the importance of the event, and practical considerations, notably whether the event rates will allow a study of reasonable size. In NI studies, the endpoint must be one for which there is a good basis for knowing the effect of the active control. The endpoint used need not necessarily be the endpoint used in the historical trials or the effectiveness endpoint claimed in labeling. Past trials, for example, with mortality endpoints could, if data were available, be the basis for estimating an effect on a composite endpoint (cardiovascular mortality, myocardial infarction, and stroke), if that were the desired endpoint for the NI study. Such a change might be sought because it would permit a smaller study or was more feasible given current event rates.

7. Are there circumstances where it may not be feasible to perform an NI study?

Unfortunately, these are many, including some where a placebo-controlled study would not be considered ethical. Some examples include the following:

- The treatment effect may be so small that the sample size required to do a non-inferiority study may not be feasible.
- There is large study-to-study variability in the treatment effect. In this case, the treatment effect may not be sufficiently reproducible to allow for the determination of a sufficiently reliable estimate of M_1.
- There is no historical evidence to determine a non-inferiority margin.
- Medical practice has changed so much (e.g., the active control is always used with additional drugs) that the effect of the active control in the historical studies is not clearly relevant to the current study.

8. In a situation where a placebo-controlled trial would be considered unethical, but a non-inferiority study cannot be performed, what are the options?

In that case it may be possible to design a superiority study that would be considered ethical. These possibilities are discussed in section III of this guidance and ICH E-10, and include the following:

- When the new drug and established treatment are pharmacologically distinct, an add-on study where the test drug and placebo are each added to the established treatment.
- A study in patients who do not respond to the established therapy. It may be possible to do a placebo-controlled trial in those patients. To establish specific effectiveness in non-responders, the study should randomize to test drug and the failed therapy and show superiority of the test drug.
- A study in patients who cannot tolerate the established effective therapy.
- A study of a population in which the effect of available therapy is not established.
- For a drug with dose-related side effects, and where a dose lower than the usual dose would be considered ethical, a dose-response study may be possible.

9. When will a single NI study be sufficient to support effectiveness?

Several sections above touch on this question, notably III.B.2, which discusses it in detail. Briefly, reliance on a single study in the NI setting is based on considerations similar to reliance on a single study in the superiority setting, with the additional consideration of the stringency of showing NI using the M_2 NI margin. Many of these factors are described in the guidance for industry on *Providing Clinical Evidence of Effectiveness for Human Drugs and Biological Products*, and include prior supportive information, such as results with pharmacologically similar agents (a very common consideration, as the NI study will often compare drugs of the same pharmacologic class), support from credible biomarker information (tumor responses, ACE inhibition,

beta blockade), and a statistically persuasive result. With respect to the latter, it is noted above that a finding of NI based on excluding a treatment difference > M_2 provides very strong evidence (generally equivalent to a $p < 0.001$ in a superiority setting) that the test treatment has an effect > 0. For all these reasons, most NI studies with outcome endpoints, if clearly successful, will be supportive as single studies. Of course, the importance of the study endpoint will influence the level of assurance needed, in a single study or multiple studies, that no more than M_2 has been lost.

APPENDIX — EXAMPLES

The following five examples derived from publicly available information (see references following examples) illustrate different aspects of the process of choosing a NI margin, of the application of a method of NI analysis, and other considerations relevant to whether it is possible to conduct and interpret the results of a NI study

Example 1(A): Determination of an NI Margin for a New Anticoagulant — Fixed Margin Approach

This example will demonstrate the following points:

- The determination of the NI margin (M_1) using the fixed margin approach
- How to select and assess the randomized trials of the active control on which to base the estimate of active comparator treatment effect.
- How to assess whether the assumption of assay sensitivity is appropriate, and whether the constancy assumption is reasonable for this drug class.
- Why it is appropriate to use a conservative choice (e.g., 95% lower bound) for estimating the treatment effect size of the active comparator, accounting for between-study variability, and considering other uncertainties in the randomized trial data.
- The use of the lower bound of the 95% confidence interval in the NI study for C-T to demonstrate non-inferiority.

SPORTIF V is an NI study that tested the novel anticoagulant ximelagatran against the active control warfarin. Warfarin is a highly effective, orally active anticoagulant that is approved in the United States for the treatment of patients with non-valvular atrial fibrillation at risk of thromboembolic complications (e.g., stroke, TIA, etc.). There are six placebo-controlled studies of warfarin involving the treatment of patients with non-valvular atrial fibrillation, all published between the years 1989 and 1993. The primary results of these studies are summarized in Table 1 and provide the basis for choosing the NI margin for SPORTIF V.

The point estimate of the event rate on warfarin compared to placebo is favorable to warfarin in each of the 6 studies. The upper bound of the 95% confidence interval of the risk ratio calculated in each study is less than one in five of the six studies, indicating a statistically demonstrated treatment effect in each of these studies. The one exception is the CAFA study. However, this study was reportedly stopped early because of favorable results published from the AFASAK and SPAF I studies (Connolly et al. 1991). Although the CAFA study was stopped early, a step that can sometimes lead to an overestimate of effect, the data from this study appear relevant in characterizing the overall evidence of effectiveness of warfarin because there is no reason to think it was stopped for early success, introducing a possible favorable bias. These placebo controlled studies of warfarin in

patients with non-valvular atrial fibrillation show a fairly consistent and reproducible effect. Based on the consistent results from the six studies, it can reasonably be assumed that were placebo to be included in a warfarin-controlled NI study involving a novel anticoagulant, warfarin would have been superior to placebo.

Table 1: Placebo-Controlled Trials of Warfarin in Non-Valvular Atrial Fibrillation

Study	Summary	Events/Patient Years		Risk Ratio (95% CI)
		Warfarin	Placebo	
AFASAK	open label. 1.2 yr follow-up	9/413 = 2.18%	21/398 = 5.28%	0.41 (0.19, 0.89)
BAATAF	open label. 2.2 yr follow-up	3/487 = 0.62%	13/435 = 2.99%	0.21 (0.06, 0.72)
EAFT	open label. 2.3 yr follow-up patients with recent TIA	21/507 = 4.14%	54/405 = 13.3%	0.31 (0.19, 0.51)
CAFA*	double blind. 1.3 yr follow-up	7/237 = 2.95%	11/241 = 4.56%	0.65 (0.26, 1.64)
SPAF I	open label. 1.3 yr follow-up	8/260 = 3.08%	20/244 = 8.20%	0.38 (0.17, 0.84)
SPINAF	double blind. 1.7 yr follow-up	9/489 = 1.84%	24/483 = 4.97%	0.37 (0.17, 0.79)

* CAFA was stopped early because of favorable results observed in other studies.

As can be seen from the summary table, most of these studies were open label. It is not clear how great a concern this should be given the reasonably objective endpoints in the study (see Table 2), but to the extent there is judgment involved, there is some possible bias. The event rate on placebo in the EAFT study was strikingly high, perhaps because the patient population in that study was different from the patient population studied in the remaining five studies in that only patients with a recent TIA or stroke were enrolled in EAFT. That would clearly increase the event rate, but in fact the risk reduction in EAFT was very similar to the four trials other than CAFA, which is relatively reassuring with respect to constancy of risk reduction in various AF populations.

Even if the historical studies are consistent, a critical consideration in deciding upon the NI margin derived from these studies is whether the constancy assumption is reasonable. The constancy assumption must consider whether the magnitude of effect of warfarin relative to placebo in the previous studies would be present in the new NI study, or whether changes in medical practice (e.g., concomitant medications, skill at reaching desired INR), or changes in the population being tested may make the effect of warfarin estimated from the previous studies not relevant to the current NI study.

To evaluate the plausibility of this constancy assumption, one might compare some features of the six placebo-controlled warfarin studies with the NI study, SPORTIF V. There is considerable heterogeneity in the demographic characteristics of these studies. While some study subject characteristics can be compared across the studies (e.g., age, race, and target INR) certain characteristics cannot be compared (e.g., concomitant medication use, race, mean blood pressure at baseline) if they are not consistently reported in the study publications. Whether these are critical to outcomes is, of course, the critical question. Table 2 indicates that for some characteristics, such as a history of stroke or TIA, there are inter-study differences. One of the important inclusion criteria in the EAFT study was that

subjects had a prior history of stroke or TIA. None of the other studies had such a requirement. Subjects enrolled into the EAFT study were thus at higher risk than subjects in the other studies, presumably leading to the higher event rates in both the warfarin and placebo arms, shown in Table 1. The higher event rates in the EAFT study may also have been influenced by the relatively long duration of follow-up or the fact that the primary endpoint definition was broader, including vascular deaths and non-fatal myocardial infarctions, which might have been less affected by coumadin, leading to a lower risk reduction. This was not in fact seen. All in all, the results are quite consistent (with the exception of CAFA), a relatively reassuring outcome.

Table 2: Demographic Variables, Clinical Characteristics, and Endpoints of Warfarin AF Studies

	AFASAK	BAATAF	CAFA	SPAF	VA	EAFT	SPORTIF V
Age years (mean)	73	69	68	65	67	71	72
Sex (%) Male	53%	75%	76%	74%	100%	59%	70%
h/o stroke or TIA (%)	6%	3%	3%	8%	0%	100%	18.3%
h/o HTN (%)	32%	51%	43%	49%	55%	43%	81%
≥65 years old & CAD (%)*	8%	10-16%	12-15%	7%	17%	7%	41%
>65 years old & DM (%)*	7-10%	14–16%	10-14%	13%	17%	12%	19%
h/o LV dysfunction (%)*	50%	24-28%	20-23%	9%	31%	8%	39%
Mean BP at BL (mm Hg)	NA	NA	NA	130/78	NA	145/84	133/77
Target INR	2.8-4.2	1.5-2.7	2-3	2-4.5	1.4-2.8	2.5-4.0	2-3
Primary endpoint	Stroke, TIA, systemic embolism	Ischemic stroke	Ischemic stroke and systemic embolism	Ischemic stroke and systemic embolism	Ischemic stroke	Vascular death, NF MI, stroke, systemic embolism	Stroke (ischemic + hemorrhagic) and systemic embolism

* = Not possible to verify whether definitions of CAD, DM, and LV dysfunction were the same in comparing the historic studies and SPORTIF V.
NA = Not available

At the time the SPORTIF V study was reviewed, concerns about whether the constancy assumption held and other factors led to the consideration of whether discounting of the effect size would be appropriate (see discussion of discounting in section IV of this guidance). We now believe the historic results are reasonably likely to be consistent with results that would be seen today so that discounting was not necessary. To calculate M_1, the relative risks in each of the six studies were combined using a random effects model to give a point estimate of 0.361 for the relative risk with a confidence interval of (0.248, 0.527). The 95% CI upper bound of 0.527 represents a 47% risk reduction, which translates into a risk increase of about 90% from not being on warfarin (1/0.527 = 1.898) (i.e., what would be seen if the test drug had no effect). Thus, M_1 (in terms of the hazard ratio favoring the control to be ruled out) is 1.898.

It was considered clinically necessary to show that the test drug preserved a substantial fraction of the warfarin effect. The clinical margin M_2 representing the largest acceptable inferiority of the test to control, was therefore set at 50% of M_1. As described in section IV of the guidance, we calculate M_2, using the log hazard risk ratios, as 1.378, 95% CI for C-T < 1.378.

In the SPORTIF V study, the point estimate of the relative risk was 1.39 and the two-sided 95% confidence interval for the relative risk was (0.91, 2.12). Thus, in this example, the non-inferiority of ximelegatran to warfarin is not demonstrated because the upper limit (2.12) is greater than M_2 (=1.378). Indeed, it does not even demonstrate that M_1 (=1.898) has been excluded.

This example illustrates the fixed margin approach and what is often called the "two 95% confidence interval approach." That is, a two-sided 95% confidence interval is used for the historical data to select M_1, and a two-sided 95% confidence interval is used to test whether M_2 has been ruled out, similar to controlling the Type 1 error of the NI study at one-sided 2.5%.

Example 1(B): Application of the Synthesis Method to the Above Example 1(A)

This example demonstrates the following:

- The critical features of the synthesis approach to demonstrating the NI of a new anticoagulant.

- The calculations and sources of statistical variability that are incorporated in the synthesis approach.

- The main differences in interpretation of the fixed margin and the synthesis approaches when applied to the same set of studies and data.

In this example, we illustrate the synthesis method using the same data as Example 1(A), which consist of six studies comparing warfarin to placebo and one NI study comparing ximelegatran to warfarin. In contrast to the fixed margin method in Example 1(A), the synthesis method does not use a separate 95% confidence interval for this historical estimate of the effect of warfarin versus placebo and for the comparison in the NI study. Rather, the synthesis method is constructed to address the questions of whether ximelegatran preserves a specified percent, in this case 50% or one-half (versus placebo), of the effect of warfarin, and whether ximelegatran would be superior to a placebo, if one had been included as a randomized treatment group in the NI study. To accomplish this goal, the synthesis method makes a comparison of the effect of ximelegatran in the NI study to historical placebo data, an indirect comparison that is not based upon a randomized current placebo group. The synthesis method combines the data from the placebo-controlled studies of warfarin with the data from the NI study in such a way that a test of hypothesis is made to demonstrate that a certain percent of the effect of warfarin is retained in the NI study. A critical point distinguishing the synthesis method from the fixed margin method is that the M_1 effect size of warfarin is not specified in advance and is not required to be fixed prior to carrying out the synthesis method. But to carry out the analysis, an assumption needs to be made regarding the placebo comparison, namely, that the difference between control drug and placebo (had there been one) in the NI trial is the same as what was seen in the historical placebo-controlled trials of warfarin. The assumption is needed because there is no randomized comparison of warfarin and placebo in the NI trial. As a point of reference, we know from the previous example, 1(A), that the warfarin effect M_1 was estimated from the historical placebo studies to be a 47% risk reduction.

In this case, the synthesis method statistically tests the null hypothesis that the inferiority of ximelegatran compared to warfarin is less than 50% or one half of the risk reduction of warfarin compared to placebo, a question that the fixed margin method does not directly address because in the fixed margin method, the placebo is only present in the historical studies and not in the NI study. We carry out this test on the log relative risk scale, so that the null hypothesis can be written as:

H_0: {log-Relative Risk of ximelegatran versus warfarin} ≥

– ½ {log-Mean Relative Risk of warfarin versus placebo}

A test of this hypothesis is performed by the expression below (the statistical test) that has the form of a quotient where the numerator is an estimate of the parameter defined in the null hypothesis by {log-Relative Risk of ximelegatran versus warfarin} + ½ {log-Mean Relative Risk of warfarin versus placebo} and the denominator is an estimate of the standard error of the numerator. In this case, the estimated log-Relative Risk of ximelegatran versus warfarin is 0.329 (log of 1.39) with a standard error of 0.216 while the estimated log-Relative Risk of warfarin versus placebo is -1.02 (log of .527) with a standard error of 0.154. The estimate of the log warfarin effect is -1.02, and the standard error of this estimate is 0.154; these estimates are combined with the NI data as if all the data were in a randomized comparison with placebo. The synthesis test statistic is calculated as:

$$\frac{0.329 + \frac{1}{2}\{-1.02\}}{\sqrt{0.216^2 + \left\{\frac{1}{2}\{0.154\}\right\}^2}} = -0.789$$

Assuming the statistic is normally distributed, it is then compared to -1.96 (for one-sided Type 1 error rate of 0.025). For this case, the value, -0.789, is not less (more negative) than -1.96, so we cannot reject the null hypothesis. Therefore, it cannot be concluded that an NI margin of 50% retention is satisfied.

To compare the fixed margin method with the synthesis method, recall that the fixed margin compares the upper or lower limits of two 95% confidence intervals, one for the NI study and one for the meta-analysis of the effect of warfarin. One might consider the fixed margin approach as conservative, as it compares to statistically "worst cases." The synthesis method does not use two such worst cases. To provide a more detailed comparison of the approaches, the fixed margin approach can be expressed as using a test statistic similar to that of the synthesis approach.

The synthesis method concludes non-inferiority if

$$\frac{0.329 + \frac{1}{2}\{-1.02\}}{\sqrt{0.216^2 + \left\{\frac{1}{2}\{0.154\}\right\}^2}} < -1.96$$

The fixed margin method concludes non-inferiority if

$$\frac{0.329 + \frac{1}{2}\{-1.02\}}{0.216 + \frac{1}{2}\{0.154\}} < -1.96$$

The critical difference between these two procedures is the form of the denominator, which expresses the standard errors of the expressions in the numerator. The synthesis standard error is always smaller than that of the fixed margin method when expressed in this manner. In most situations, the synthesis is therefore statistically more efficient (and would require a smaller sample size) than the fixed margin approach. Of course, the approach can be considered useful and valid only if the assumptions of the synthesis method can be considered satisfied. This is not always possible, generally because of concerns about constancy, that is, whether the historical differences from placebo would accurately describe the current differences from placebo.

The two procedures also cannot be directly compared because they have other differences that make their comparison problematic, notably the differences in how the statistical error rates, or Type 1 errors, are calculated and interpreted. The synthesis method, because of the way it makes the comparisons with a placebo, gives equal weight to the variance (or variability of the outcome data) in this historical estimate and the variance of the data obtained from the randomized comparison of the test drug and active comparator in the NI study. When the historical database is very large relative to the NI database, combining the historical data and NI together may suggest greater precision in the overall assessment of the NI study than is warranted given the fact that the placebo comparisons were from studies conducted in a different population, usually at a different time. In contrast, the fixed margin method controls a Type 1 error rate within the NI study that is conditioned on the pre-specified fixed NI margin, separately estimated from the historical active comparator data. The synthesis test method also does not estimate a fixed NI margin to be excluded (i.e., one depending only on the prior placebo-controlled data for the active comparator).

A general principle expressed in this guidance is the need to be conservative in the selection of the margin M_1 because that margin is critical to establishing that a test drug is effective in an NI study design. The M_1 margin is usually chosen conservatively because of the uncertainties associated with the validity of assumptions in an NI study and the reliance on historical active control comparisons. As noted, the fixed margin approach can be considered conservative in that several worst case situations (lower bounds of 95% confidence intervals) are used, one evaluating the historical evidence and another in the NI comparison. We recommend use of this conservative fixed margin approach to selecting the M_1 margin and to demonstrating in the NI study that the M_1 margin is excluded at the acceptable Type 1 error. The synthesis method, on the other hand, as described above, is less conservative. But this is reasonable, given that M_2 is considerably smaller (a more demanding margin) and that the presence of a control drug effect has been well established by ruling out loss of M_1 using the fixed margin approach. We therefore believe the NI study

should utilize a fixed margin approach to ruling out loss of M_1 but can use the synthesis method to establish that loss of effect greater than the clinically relevant margin M_2 has been ruled out.

Example 2: The Determination of a Non-Inferiority Margin for Complicated Urinary Tract Infection (cUTI) — Fixed Margin Approach

This example will illustrate the following points:

- The use of the absolute difference in cure rates as the metric of treatment effect.
- The determination of a non-inferiority margin when there are no randomized active comparator placebo-controlled studies available for the indication of interest (in this case, cUTI).
- Estimating the placebo response rate in cUTI based upon data from uncomplicated urinary tract infections (a generally less severe form of urinary tract infection leading to a high, therefore conservative, estimate).
- The importance of seeking out all relevant studies for the margin determination and incorporating the limitations of the studies, the analyses, and the resulting estimates in the consideration of the resulting estimate of the non-inferiority margin.
- This approach (i.e., relying on data other than controlled trials of the active control) is credible only when the effect size is large, given its limitations.

The following steps were used to estimate the effectiveness of the active control.

1. Evaluation of the placebo response rate in uncomplicated urinary tract infection (uUTI)
2. Evaluation of outcomes in patients receiving inadequate or inappropriate therapy for complicated urinary tract infection (cUTI)/acute pyelonephritis (AP)
3. Evaluation of the active comparator's response rate (levofloxacin, in this case) for cUTI.

Step 1: Placebo Response Rate for Uncomplicated Urinary Tract Infection (uUTI)

Although there were no placebo-controlled complicated UTI studies available, three placebo-controlled studies in women with uncomplicated UTI were identified. Among these three studies there were differences in the duration of study drug, endpoints assessed, and the diagnostic criteria for significant bacteriuria. There were no placebo-controlled trials identified in men with UTI without significant co-morbid conditions, and the pathophysiology and natural history of UTI are different in men and women. It would be expected that placebo response rates would therefore be high in such studies compared to the untreated rate in cUTI and represent a conservative (high) estimate of the spontaneous cure rate in cUTI.

Microbiological eradication rate is generally used as the primary endpoint for UTI studies. In the three placebo-controlled studies identified for UTI, the bacteriological response rates were 95/227(42%) for the combined 8-10 and 35-49 days (Ferry et al.), 9/27(33%) at day 3 (Christiaens et al.), and 8/18(44%) in 1 week (Dubi et al.). The bacteriologic criteria for entry used in the Ferry study were $\geq 10^3$ CFU/ml for primary pathogens, whereas $\geq 10^4$ CFU/ml was used for the Christiaens study. Because a count of $\geq 10^5$ CFU/ml is more

typically used as diagnostic criteria for a uropathogen, the studies could overestimate the placebo response rates by including patients whose colony counts would not cause them to be considered infected. The results are summarized in the following table.

Table 3: Historical Placebo Data from Published uUTI Studies

Author	Type of UTI	Placebo	95% CI[1]
Ferry et al.	uUTI	95/227 (42%)	(35.4 %, 48.6%)
Christiaens et al.	Acute uUTI	9/27 (33%)	(16.5%, 54.0%)
Dubi et al.	uUTI	8/18 (44%)	(21.5%, 69.2%)

[1]Exact Confidence Intervals

Because of the unequal study population sizes, a weighted analysis is needed. The weighted non-iterative method for random effects model using logit of the event rates described by DerSimonian and Laird was used to obtain the estimate and its 95% CI; the weighted estimate is 41.2% with 95% CI of (35.5%, 47.2%).

Step 2: Outcomes Subsequent to Inadequate or Inappropriate Antibacterial Therapy for Complicated Urinary Tract Infection (cUTI)/AP

Three studies were identified in which some patients were treated with an antimicrobial drug to which the bacteria causing their UTI were resistant (inadequate therapy). Eradication rates for pathogens resistant to the antimicrobial drug may be considered as another way to estimate the placebo effect in cUTI/AP. It should be noted, however, that the use of data from inadequate therapy may result in an estimate that is higher than a true placebo, once again a conservative estimate of effect, because even "inadequate" therapy may have some effect on the patient's infection.

Table 4: Eradication Rates in Patients Receiving Inadequate Therapy

Author	Type of UTI	Eradication Rates	95% CI[1]
Allais et al.	cUTI/AP	12/23 (52.2%)	(30.6%, 73.2%)
Fang et al.	cUTI/AP	4/28 (14.3%)	(4.0%, 32.7%)
Talan et al.	AP	7/14 (50.0%)	(23.0%, 77.0%)

[1]Exact Confidence Intervals

The data from the historical studies in Table 4 were combined to obtain a weighted estimate of the inadequate therapy eradication rate and its corresponding two-sided 95% CI. The weighted estimate using the DerSimonian and Laird approach (random effect model) is 36.8% with 95% CI of (15.4%, 64.9%).

Step 3: Active Comparator's Eradication Rate for Complicated UTI (cUTI)

To assess the eradication rates for the active comparator, levofloxacin, four cUTI studies were considered, including two published studies and two studies submitted to the Agency (Study A and Study B) that involved men and women ≥18 years old. The two studies from

the medical literature had limitations. In the Peng study, the microbiological eradication rate was evaluated on Day 5, while antibiotic therapy was still ongoing. This could have falsely elevated the response rate. The Klimberg study was an open-label study, and was excluded from the analysis because of concern about potential bias.

The other two studies, Study A and Study B, were blinded controlled studies using levofloxacin for the treatment of cUTI. In Study A, the microbiological eradication rate for levofloxacin was 84.2% (154/183). In Study B, the microbiological eradication rate for levofloxacin was 78.2% (252/321). The levofloxacin eradication rates for the Peng study and Studies A and B are shown in Table 5. The weighted estimate of eradication rates using the DerSimonian and Laird approach is 81.6% with 95% CI of (75.8%, 86.3%).

Table 5: Historical Levofloxacin Data from Published cUTI Studies

Author	Type of UTI	Levofloxacin Microbiological Eradication Rate	95% CI[1]
Peng et al.	cUTI	18/20 (90%)	(68.3%, 98.8%)
Study A	cUTI and AP	154/183 (84.2%)	(78.0%, 89.1%)
Study B	cUTI and AP	252/321 (78.2%)	(73.6%, 82.9%)

[1]Exact confidence intervals

Step 4: Estimated Non-Inferiority Margin for Complicated UTI (cUTI) Using Levofloxacin as the Active Comparator

The placebo eradication rate is estimated from the upper bound of the two-sided 95% CI for the placebo eradication rate in uUTI (47%) and this estimate is supported by evidence based on outcomes subsequent to inadequate or inappropriate therapy in cUTI (65%). The estimated levofloxacin cure rate for sensitive organisms is 76% (using the lower bound of the 95% CI for the weighted levofloxacin response rate). Using the placebo eradication rate for uUTI, the historical treatment effect can be calculated as 29% (=76%-47%). The treatment effect based on outcomes following inadequate antibacterial therapy can be calculated as 11% (=76%-65%), providing supportive evidence.

Major Limitations in This Example:

Apart from the lack of a direct comparison of active control and placebo in cUTI, there were various uncertainties in the historical estimates described above because of problems with data quality, study design, population size, prognostic factors, and differences in the timing of the microbiological endpoint assessments. On the other hand, the placebo eradication rate was estimated based on placebo-controlled clinical studies assessing the antibacterial treatment in a population (female subjects with uUTI) that would almost certainly give an overestimate of the spontaneous or placebo eradication rate in cUTI, leading to a conservative (low) estimate of the effect of the active control.

Discounting and Preservation of the Levofloxacin Treatment Effect:

The various limitations and uncertainties in the historical data led to discounting of the calculated treatment effect of 29%. Thus, the active control treatment effect over placebo (M_1) was estimated as 14.5% based on a 50% discounting. For a serious illness, a substantial portion (at least 50% or more) of M_1 should be preserved. Accordingly, an NI margin of 7% was specified as M_2 based on clinical judgment.

1991
1992 **Example 3: Aspirin to Prevent Death or Death/MI After Myocardial Infarction**
1993
1994 This example demonstrates the following:
1995
1996 • When it may not be possible to determine the NI margin because of the limitations of the
1997 data available.
1998
1999 By 1993, the effect of aspirin in preventing death after myocardial infarction had been
2000 studied in six large randomized placebo-controlled clinical trials. A seventh trial, ISIS-2,
2001 gave the drug during the first day after the AMI and is not included because it addressed a
2002 different question. The results are summarized and presented in chronological order in Table
2003 6.
2004
2005 **Table 6. Results of six placebo-controlled randomized studies (listed in chronological order) of the effect**
2006 **of aspirin in preventing death after myocardial infarction**

Study	Year published	Aspirin		Placebo		Relative Risk (95% CI)
		N	Death rate	N	Death rate	
MRC-1	1974	615	8.0%	624	10.7%	0.74 (0.52, 1.05)
CDP	1976	758	5.8%	771	8.3%	0.70 (0.48, 1.01)
MRC-2	1979	832	12.2%	850	14.8%	0.83 (0.65, 1.05)
GASP	1978	317	10.1%	309	12.3%	0.82 (0.53, 1.28)
PARIS	1980	810	10.5%	406	12.8%	0.82 (0.59, 1.13)
AMIS	1980	2267	10.9%	2257	9.7%	1.12 (0.94, 1.33)

2007
2008 The results suggest:
2009
2010 (1) The effect of aspirin on mortality as measured by the relative risk seems to
2011 attenuate over the time the studies were conducted.
2012 (2) The largest trial, AMIS, showed a numerically adverse effect of aspirin.
2013
2014 The relative risk in the AMIS study is significantly different from the mean relative risk in
2015 the remaining studies ($p \leq 0.005$). The validity of pooling the results of AMIS with those of
2016 the remaining studies is therefore a concern. It would be invalid to exclude AMIS from the
2017 meta-analyses because its effect differed from the effect in the remaining studies, unless there
2018 were adequate clinical or scientific reasons for such exclusion. At a minimum, any meta-
2019 analysis of all studies would need to reflect this heterogeneity by using a random-effect
2020 analysis.
2021
2022 Although a fixed effect analysis of the six studies gives a point estimate of 0.91 (95% CI 0.82
2023 to 1.02), the random-effects analysis gives a point estimate of 0.86 with 95% confidence
2024 interval (0.69, 1.08). The effect of aspirin on prevention of death after myocardial infarction
2025 in these historical studies is thus inconclusive (i.e., the upper bound of the 95% CI for effect
2026 is > 1.0). Therefore, it would be difficult, indeed not really possible, to select aspirin as the

active control for evaluating the mortality effect of a test drug in a non-inferiority trial. Apart from this calculation, it seems difficult to accept an NI endpoint that is not supported by the largest of the six trials.

The same six studies can also be examined for the combined endpoint of death plus AMI in patients with recent AMI. This endpoint reflects the current physician-directed claim for aspirin based on the positive finding in two studies (MRC-2, PARIS).

Table 7. Results of six placebo-controlled randomized studies of the effect of aspirin in secondary prevention of death or MI after myocardial infarction

Study	Year published	Aspirin		Placebo		Relative Risk (95% CI)
		N	Event rate	N	Event rate	
MRC-1	1974	615	9.9%	624	13.1%	0.75 (0.55, 1.03)
CDP	1976	758	9.5%	771	12.5%	0.76 (0.57, 1.02)
MRC-2	1979	832	16.0%	850	22.2%	0.72 (0.59, 0.88)
GASP	1978	317	13.6%	309	17.5%	0.78 (0.54, 1.12)
PARIS	1980	810	17.4%	406	22.7%	0.77 (0.61, 0.97)
AMIS	1980	2267	18.6%	2257	19.2%	0.97 (0.86, 1.09)

***the event rate of either group needs further verification from each article**

The results indicate that the effect of aspirin on death or MI after myocardial infarction is small to absent in the latest trial (AMIS). Random-effect analyses give, depending on the specific analysis, point estimates of the relative risk of 0.81-0.85, with 95% CI upper bounds of 0.96-1.02. The NI margin based on these six studies ranges from 4% to zero (without reducing it further to represent M_2) is so small that a trial to rule out loss at this effect would be unrealistically large. Again, as with the mortality endpoint, it would be troubling even to consider an NI approach when the largest and most recent trial showed no significant effect.

Example 4: Xeloda to Treat Metastatic Colorectal Cancer - the Synthesis Method

This example of Xeloda for first-line treatment of metastatic colorectal cancer illustrates:

- The use of the synthesis method to demonstrate a loss of no more than 50% of the historical control treatment's effect and a relaxation of this criterion when two NI studies are available.

- The use of supportive endpoints in the decision making process.

- The use of a conservative estimate of the control treatment effect size, because a subset of the available studies to estimate the margin was selected and the effect was measured relative to a previous standard of care instead of placebo.

The U.S. regulatory standard for first-line treatment of metastatic colorectal cancer, the use sought for Xeloda, is the demonstration of improvement in overall survival. Two separate clinical trials, each using an NI study design, compared Xeloda to a Mayo Clinic regimen of 5-fluorouracil with leucovorin (5-FU+LV), the standard of care at the time. Xeloda is an oral fluoropyrimidine, while 5-fluorouracil (5-FU) is an infusional fluoropyrimidine

By itself, bolus 5-FU had not demonstrated a survival advantage in first-line metastatic colorectal cancer. But with the addition of leucovorin to bolus 5-FU, the combination had demonstrated improved survival. A systematic evaluation of approximately 30 studies that investigated the effect of adding leucovorin to a regimen of 5-FU identified ten clinical trials that compared a regimen of 5-FU+LV similar to the Mayo clinic regimen to 5-FU alone, thereby providing a measure of the effect of LV added to 5-FU, a conservative estimate of the overall effect of 5-FU+LV, as it is likely 5-FU has some effect.

Table 8 summarizes the overall survival results, using the metric "log hazard ratio" for the ten studies identified that addressed the comparison of interest.

Table 8: Selected studies comparing 5FU to 5-FU+LV

Study	Hazard Ratio[1]	Log Hazard Ratio[1]	Standard Error
Historical Study 1	1.35	.301	.232
Historical Study 2	1.26	.235	.188
Historical Study 3	0.78	-.253	.171
Historical Study 4	1.15	.143	.153
Historical Study 5	1.39	.329	.185
Historical Study 6	1.35	.300	.184
Historical Study 7	1.38	.324	.166
Historical Study 8	1.34	.294	.126
Historical Study 9	1.03	.0296	.165
Historical Study 10	1.95	.670	.172

[1] All log hazard ratios are 5-FU/5-FU+LV

A random effects model applied to the survival results of these ten studies yielded the historical estimate of the 5-FU versus 5-FU+LV survival comparison of log hazard ratio of 1.264 with a 95% confidence interval of (1.09, 1.46) and a log hazard ratio of 0.234. The NI margin is therefore 1.09 for a fixed margin approach ruling out M_1.

A summary of the survival results based on the intent-to-treat populations for each of the two Xeloda NI trials is presented in Table 9. Study 2 rules out M_1 using a fixed margin approach, but Study 1 does not.

Table 9: Summary of the survival results

Study	Hazard Ratio[1]	Log Hazard Ratio[1]	Standard Error	95% CI for the Hazard Ratio[1]
NI Study 1	1.00	-0.0036	0.0868	(0.84, 1.18)
NI Study 2	0.92	-0.0844	0.0867	(0.78, 1.09)

[1] Hazard ratios and log hazard ratios are Xeloda/5-FU+LV

The clinical choice of how much of the effect on survival of 5-FU+LV should be shown not to be lost by Xeloda was determined to be 50%. The synthesis approach was used to analyze whether the NI criteria of 50% loss was met. This synthesis approach to the non-inferiority test procedure for each study combines the results of each NI study with the results from the random effects meta-analysis into a normalized test statistic.

Based on this NI synthesis test procedure, NI Study 1 failed to demonstrate that Xeloda retained at least 50% of the historical effect of 5-FU+LV versus 5-FU on overall survival, but NI study 2 did demonstrate such an effect. It was then decided to determine what percent retention might be satisfied by the data in a statistically persuasive way. By adapting the synthesis test procedure for retention of an arbitrary percent of the 5-FU+LV historical effect, it was determined that NI Study 1 demonstrated that Xeloda lost no more than 90% of the historical effect of 5-FU+LV on overall survival and that NI Study 2 demonstrated no more than a 39% loss of the historical effect.

The evidence of effectiveness of Xeloda was supported by the observation that the tumor response rates were statistically significantly greater for the Xeloda arm and the fact that Xeloda and 5-FU were structurally and pharmacologically very similar.

REFERENCES - EXAMPLES

Example 1(A)

The Boston Area Anticoagulation Trial for Atrial Fibrillation Investigators (1990). "The Effect of Low-Dose Warfarin on the Risk of Stroke in Patients with Nonrheumatic Atrial Fibrillation." *New Engl J Med* 323, 1505-1511.

Connolly, S.J., Laupacis, A., Gent, M., Roberts, R.S., Cairns, J.A., Joyner, C. (1991). "Canadian Atrial Fibrillation Anticoagulation (CAFA) Study." *J Am Coll Cardiol* 18, 349-355.

EAFT (European Atrial Fibrillation Trial) Study Group (1993). "Secondary Prevention in Non-Rheumatic Atrial Fibrillation After Transient Ischemic Attack or Minor Stroke." *Lancet* 342, 1255-1262.

Ezekowitz, M.D., Bridgers, S.L., James, K.E., Carliner, N.H., et al. (1992). "Warfarin in the Prevention of Stroke Associated with Nonrheumatic Atrial Fibrillation." *N Engl J Med* 327, 1406-1412.

Food and Drug Administration, Dockets home page. Available at: http://www.fda.gov/ohrms/dockets/AC/04/briefing/2004-4069B1_07_FDA-Backgrounder-C-R-stat%20Review.pdf.

Halperin, J.L., Executive Steering Committee, SPORTIF III and V Study Investigators (2003). "Ximelagatran Compared with Warfarin for Prevention of Thromboembolism in Patients with Nonvalvular Atrial Fibrillation: Rationale, Objectives, and Design of a Pair of Clinical Studies and Baseline Patient Characteristics (SPORTIF III and V)." *Am Heart J* 146, 431-8.

Jackson, K., Gersh, B.J., Stockbridge, N., Fleiming, T.R., Temple, R., Califf, R.M., Connolly, S.J., Wallentin, L., Granger, C.B. (2005). Participants in the Duke Clinical Research Institute/American Heart Journal Expert Meeting on Antithrombotic Drug Development for Atrial Fibrillation (2008). "Antithrombotic Drug Development for Atrial Fibrillation: Proceedings." Washington, D.C., July 25-27, 2005. *American Heart Journal* 155, 829-839.

Petersen, P., Boysen, G., Godtfredsen, J., Andersen, E.D., Andersen, B. (1989). "Placebo-Controlled, Randomised Trial of Warfarin and Aspirin for Prevention of Thromboembolic Complications in Chronic Atrial Fibrillation." *The Lancet* 338, 175-179.

Stroke Prevention in Atrial Fibrillation Investigators (1991). "Stroke Prevention in Atrial Fibrillation Study: Final Results." *Circulation* 84, 527-539.

Example 1(B) Refer to "General Reference" Section for synthesis methods.

Example 2

Allais, J.M., Preheim, L.C., Cuevas, T.A., Roccaforte, J.S., Mellencamp, M.A., Bittner, M.J. (1988). "Randomized, Double-Blind Comparison of Ciprofloxacin and Trimethoprim Sulfamethoxazole for Complicated Urinary Tract Infections." *Antimicrob Agents Chemother*. 32(9), 1327-30.

Christiaens, T.C., De Meyere, M., Verschraegen, G., et al (2002). "Randomised Controlled Trial of Nitrofurantoin Versus Placebo in the Treatment of Uncomplicated Urinary Tract Infection in Adult Women." *Br J Gen Pract*. 52(482), 729-34.

DerSimonian, R., Laird, N. (1986), "Meta-Analysis in Clinical Trials," *Controlled Clinical Trials.* 7, 177-188.

Dubi, J., Chappuis, P., Darioli, R. (1982). "Treatment of Urinary Infection with a Single Dose of Co-trimoxazole Compared with a Single Dose of Amoxicillin and a Placebo." *Schweiz Med Wochenschr.* 12(3), 90–92.

Fang, G.D., Brennen, C., Wagener, M. et al (1991). "Use of Ciprofloxacin Versus Use of Aminoglycosides for Therapy of Complicated Urinary Tract Infection: Prospective, Randomized Clinical and Pharmacokinetic Study." *Antimicrob Agents Chemother*. 35(9), 1849-55.

Ferry, S.A., Holm, S.E., Stenlund, H., Lundholm, R., Monsen, T.J. (2004). "The Natural Course of Uncomplicated Lower Urinary Tract Infection in Women Illustrated by a Randomized Placebo-Controlled Study." *Scan J Infect Dis.* 36, 296-301.

Ferry, S.A., Holm, S.E., Stenlund, H., Lundholm, R., Monsen, T.J. (2007). "Clinical and Bacteriological Outcome of Different Doses and Duration of Pivmecillinam Compared with Placebo Therapy of Uncomplicated Lower Urinary Tract Infection in Women: The LUTIW Project." *Scan J of Primary Health Care*. 25(1), 49-57.

Klimberg, I.W., Cox, C.E. 2nd, Fowler, C.L., King, W., Kim, S.S., Callery-D'Amico, S. (1998). "A Controlled Trial of Levofloxacin and Lomefloxacin in the Treatment of Complicated Urinary Tract Infection." *Urology*. 51(4), 610-5.

Peng, M.Y. (1999). "Randomized, Double-Blind, Comparative Study of Levofloxacin and Ofloxacin in the Treatment of Complicated Urinary Tract Infections." *J Microbiol Immunol Infect*. 32(1), 33-9.

Talan, D.A., Stamm, W.E., Hooton, T.M. et al (2000). "Comparison of Ciprofloxacin (7 Days) and Trimethoprim-Sulfamethoxazole (14 Days) for Acute Uncomplicated Pyelonephritis in Women." *JAMA. 2*83(12), 1583-1590.

Example 3

Aspirin Myocardial Infarction Study Research Group (1980). "A Randomized Controlled Trial of Aspirin in Persons Recovered from Myocardial Infarction." *JAMA* 243, 661-669.

Breddin, K., Loew, D., Lechner, K., Uberia, E.W. (1979). "Secondary Prevention of Myocardial Infarction. Comparison of Acetylsalicylic Acid, Phenprocoumon and Placebo. A Multicenter Two-Year Prospective Study." *Thrombosis and Haemostasis* 41, 225-236.

Coronary Drug Project Group (1976). "Aspirin in Coronary Heart Disease." *Journal of Chronic Disease* 29, 625-642.

Elwood, P.C., Cochrane, A.L., Burr, M.L., Sweetnam, P.M., Williams, G., Welsby, E., Hughes, S.J., Renton, R. (1974). "A Randomized Controlled Trial of Acetyl Salicylic Acid in the Secondary Prevention of Mortality from Myocardial Infarction." *British Medical Journal* 1, 436-440.

Elwood, P.C., Sweetnam, P.M. (1979). "Aspirin and Secondary Mortality After Myocardial Infarction." *Lancet* ii, 1313-1215.

Fleiss, J.L. (1993). "The Statistical Basis of Meta-Analysis." *Statistical Methods in Medical Research* 2,121-145.

ISIS-2 Collaborative Group (1988). "Randomized Trial of Intravenous Streptokinase, Oral Aspirin, Both, or Neither Among 17187 Cases of Suspected Acute Myocardial Infarction: ISIS-2." *Lancet* 2, 349-360.

Persantine-Aspirin Reinfarction Study Research Group (1980). "Persantine and Aspirin in Coronary Heart Disease." *Circulation* 62, 449-461.

Example 4

FDA Guidance for Industry: Oncologic Drugs Advisory Committee Discussion on FDA Requirements for the Approval of New Drugs for Treatment of Colon and Rectal Cancer.

FDA Medical-Statistical review for Xeloda (NDA 20-896) dated April 23, 2001. (http://www.fda.gov/cder/foi/nda/2001/20896s6_Xeloda_Medr_Statr_P1.pdf).

GENERAL REFERENCES

Blackwelder, W.C. (1982). "Proving the Null Hypothesis in Clinical Trials." *Controlled Clinical Trials* 3, 345-353.

Blackwelder, W.C. (2002). "Showing a Treatment is Good Because it is Not Bad: When Does "Noninferiority" Imply Effectiveness?" *Control Clinical Trials* 23, 52–54.

Brittain, E., Lin, D. (2005). "A Comparison of Intent-to-Treat and Per Protocol Results in Antibiotic Non-Inferiority Trials." *Statistics in Medicine* 24, 1-10.

Brown, D., Day, S. (2007). Reply. *Statistics in Medicine* 26, 234-236.

CBER/FDA Memorandum (1999). Summary of CBER Considerations on Selected Aspects of Active Controlled Trial Design and Analysis for the Evaluation of Thrombolytics in Acute MI, June 1999.

Committee for Proprietary Medicinal Products (CPMP) (2000). Points to Consider on Switching Between Superiority and Non-Inferiority. http://www.emea.europa.eu/pdfs/human/ewp/048299en.pdf.

Committee for Medicinal Products for Human Use (CHMP) (2006). "Guideline on the Choice of the Non-Inferiority Margin." *Statistics in Medicine* 25, 1628–1638.

Chow, S.C., Shao, J. (2006). "On Non-Inferiority Margin and Statistical Tests in Active Control Trial." *Statistics in Medicine* 25, 1101–1113.

D'Agostino, R.B., Massaro, J.M., Sullivan, L. (2003). "Non-Inferiority Trials: Design Concepts and Issues – the Encounters of Academic Consultants in Statistics." *Statistics in Medicine* 22, 169–186.

D'Agostino, R.B., Campbell, M., Greenhouse, J. (2005). "Non-Inferiority Trials: Continued Advancements in Concepts and Methodology." *Statistics in Medicine* 25, 1097-1099.

DerSimonian, R., Laird, N. (1986). "Meta-Analysis in Clinical Trials." *Controlled Clinical Trials* 7, 177-188.

Ellenberg, S.S., Temple, R. (2000). "Placebo-Controlled Trials and Active-Control Trials in the Evaluation of New Treatments - Part 2: Practical Issues and Specific Cases." *Annals of Internal Medicine* 133, 464-470.

Fisher, L.D., Gent, M., Büller, H.R. (2001). "Active-Control Trials: How Would a New Agent Compare with Placebo? A Method Illustrated with Clopidogrel, Aspirin, and Placebo." *American Heart Journal* 141: 26-32.

Fleming, T.R. (1987). "Treatment Evaluation in Active Control Studies." *Cancer Treatment Reports* 71, 1061-1064.

Fleming, T.R. (2000). "Design and Interpretation of Equivalence Trials." *American Heart Journal* 139, S171-S176.

Follmann, D.A., Proschan, M.A. (1999). "Validity Inference in Random-Effects Meta-Analysis." *Biometrics* 55, 732-737.

Freemantle, J., Cleland, J. Young, P., Mason, J., Harrison, J. (1999). "B Blockade After Myocardial Infarction: Systematic Review and Meta Regression Analysis." *British Medical Journal* 318, 1730-1737.

Gould, A.L. (1991). "Another View of Active-Controlled Trials." *Controlled Clinical Trials* 12, 474-485.

Holmgren, E.B. (1999). "Establishing Equivalence by Showing That a Prespecified Percentage of the Effect of the Active Control Over Placebo is Maintained." *Journal of Biopharmaceutical Statistics* 9(4), 651-659.

Hasselblad, V., Kong, D.F. (2001). "Statistical Methods for Comparison to Placebo in Active-Control Trials." *Drug Information Journal* 35, 435-449.

Hauschke, D. (2001). "Choice of Delta: A Special Case." *Drug Information Journal* 35, 875-879.

Hauschke, D., Hothorn, L.A. (2007). Letter to the Editor: An Introductory Note to the CHMP Guidelines: Choice of the Non-Inferiority Margin and Data Monitoring Committees. *Statistics in Medicine* 26, 230-233.

Hauschke, D., Pigeot, I. (2005). "Establishing Efficacy of a New Experimental Treatment in the 'Gold Standard' Design (with discussions)." *Biometrical Journal* 47, 782-798.

Holmgren, E.B. (1999). "Establishing Equivalence by Showing that a Prespecified Percentage of the Effect of the Active Control Over Placebo is Maintained." *Journal of Biopharmaceutical Statistics* 9, 651-659.

Hung, H.M.J., Wang, S.J., Tsong, Y., Lawrence, J., O'Neill, R.T. (2003). "Some Fundamental Issues with Non-Inferiority Testing in Active Controlled Clinical Trials." *Statistics in Medicine* 22, 213-225.

Hung, H.M.J., Wang, S.J., O'Neill, R.T. (2005). "A Regulatory Perspective on Choice of Margin and Statistical Inference Issue in Non-Inferiority Trials." *Biometrical Journal* 47, 28-36.

Hung, H.M.J., Wang, S.J., O'Neill, R.T. (2008). "Non-Inferiority Trial." *Wiley Encyclopedia of Clinical Trials*. Wiley, New York.

Hung, H.M.J., Wang, S.J., O'Neill, R.T. (2007). "Issues with Statistical Risks for Testing Methods in Noninferiority Trial Without a Placebo Arm." *Journal of Biopharmaceutical Statistics* 17, 201-213.

International Conference on Harmonization: *Statistical Principles for Clinical Trials* (ICH E-9), Food and Drug Administration, DHHS, 1998.

International Conference on Harmonization: *Choice of Control Group and Related Design and Conduct Issues in Clinical Trials* (ICH E-10), Food and Drug Administration, DHHS, July 2000.

Jones, B., Jarvis, P., Lewis, J.A., Ebbutt AF (1996). "Trials to Assess Equivalence: the Importance of Rigorous Methods." *British Medical Journal* 313, 36-39.

Julious, S.A., Wang, S.J. (2008). "How Biased are Indirect Comparisons Particularly When Comparisons Are Made Over Time in Controlled Trials?" *Drug Information Journal* 42, 625-633.

Koch, A., Röhmel, J. (2004). "Hypothesis Testing in the Gold Standard Design for Proving the Efficacy of an Experimental Treatment Relative to Placebo and a Reference." *Journal of Biopharmaceutical Statistics* 14, 315-325.

Kaul, S., Diamond, G.A. (2006). "Good Enough: A Primer on the Analysis and Interpretation of Non-Inferiority Trials." *Annals of Internal Medicine* 145, 62-69.

Lange, S., Freitag, G. (2005). "Choice of Delta: Requirements and Reality – Results of a Systematic Review." *Biometrical Journal* 47; 12-27.

Laster, L.L., Johnson, M.F., Kotler, M.L. (2006). "Non-Inferiority Trials: the '*at least as good as*' Criterion with Dichotomous Data." *Statistics in Medicine* 25, 1115-1130.

Lawrence, J. (2005). "Some Remarks About the Analysis of Active Control Studies." *Biometrical Journal* 47, 616-622.

Ng, T.H. (1993). "A Specification of Treatment Difference in the Design of Clinical Trials with Active Controls." *Drug Information Journal* 27, 705-719.

Ng, T.H. (2001). "Choice of Delta in Equivalence Testing." *Drug Information Journal* 35, 1517-1527.

Ng, T.H. (2008). "Noninferiority Hypotheses and Choice of Noninferiority Margin." *Statistics in Medicine* 27, 5392-5406.

Pledger, G., Hall, D.B. (1990). "Active Control Equivalence Studies: Do They Address the Efficacy Issue?" *Statistical Issues in Drug Research and Development*, Marcel Dekker, New York, 226-238.

Röhmel, J. (1998). "Therapeutic Equivalence Investigations: Statistical Considerations." *Statistics in Medicine* 17, 1703-1714.

Rothmann, M., Li, N., Chen, G., Chi, G.Y.H., Temple, R.T., Tsou, H.H. (2003). "Non-Inferiority Methods for Mortality Trials." *Statistics in Medicine* 22, 239-264.

Rothmann, M. (2005). "Type I Error Probabilities Based on Design-Stage Strategies with Applications to Noninferiority Trials." *J. of Biopharmaceutical Statistics* 15; 109-127.

Sanchez, M.M., Chen, X. (2006). "Choosing the Analysis Population in Non-Inferiority Studies: Per Protocol or Intent-to-Treat." *Statistics in Medicine* 25, 1169-1181.

Sheng, D., Kim, M.Y. (2006). "The Effects of Non-Compliance on Intent-to-Treat Analysis of Equivalence Trials." *Statistics in Medicine* 25, 1183-1190.

Siegel, J.P. (2000). "Equivalence and Noninferiority Trials." *American Heart Journal* 139: S166-S170.

Simon, R. (1999). "Bayesian Design and Analysis of Active Control Clinical Trials." *Biometrics* 55, 484-487.

Snapinn, S.M. (2004). "Alternatives for Discounting in the Analysis of Noninferiority Trials." *Journal of Biopharmaceutical Statistics* 14, 263-273.

Snapinn, S.M., Jiang, Q. (2008). "Controlling the Type I Error Rate in Non-Inferiority Trials." *Statistics in Medicine* 27**,** 371–381.

Temple, R. (1987). "Difficulties in Evaluating Positive Control Trials." *Proceedings of the Biopharmaceutical Section of American Statistical Association*, 1-7.

Temple R. (1996). "Problems in Interpreting Active Control Equivalence Trials." *Accountability in Research* 4: 267-275.

Temple, R., Ellenberg, S.S. (2000). "Placebo-Controlled Trials and Active-Control Trials in the Evaluation of New Treatments - Part 1: Ethical and Scientific Issues." *Annals of Internal Medicine* 133, 455-463.

Wang, S.J., Hung, H.M.J., Tsong, Y. (2002). "Utility and Pitfalls of Some Statistical Methods in Active Controlled Clinical Trials." *Controlled Clinical Trials* 23, 15-28.

Wang, S.J., Hung, H.M.J. (2003). "Assessment of Treatment Efficacy in Non-Inferiority Trials." *Controlled Clinical Trials* 24, 147-155.

Wang S.J., Hung H.M.J. (2003). "TACT Method for Non-Inferiority Testing in Active Controlled Trials." *Statistics in Medicine* 22; 227-238.

Wang, S.J., Hung, H.M.J., Tsong, Y. (2003). "Non-Inferiority Analysis in Active Controlled Clinical Trials." *Encyclopedia of Biopharmaceutical Statistics, 2nd Edition*. Marcel Dekker, New York.

Wiens, B. (2002). "Choosing an Equivalence Limit for Non-Inferiority or Equivalence Studies." *Controlled Clinical Trials* 23, 2-14.

Wiens, B. (2006). "Randomization as a Basis for Inference in Noninferiority Trials." *Pharmaceutical Statistics* 5, 265-271.